Mostly Harmless Econometrics: An Empiricist's Companion

Joshua D. Angrist
Massachusetts Institute of Technology

Jörn-Steffen Pischke
The London School of Economics

March 2008

Contents

List of Figures

List of Tables

Preface

The universe of econometrics is constantly expanding. Econometric methods and practice have advanced greatly as a result, but the modern menu of econometric methods can seem confusing, even to an experienced number-cruncher. Luckily, not everything on the menu is equally valuable or important. Some of the more exotic items are needlessly complex and may even be harmful. On the plus side, the core methods of applied econometrics remain largely unchanged, while the interpretation of basic tools has become more nuanced and sophisticated. Our *Companion* is an empiricist's guide to the econometric essentials . . . *Mostly Harmless Econometrics.*

The most important items in an applied econometrician's toolkit are:

1. Regression models designed to control for variables that may mask the causal effects of interest;
2. Instrumental variables methods for the analysis of real and natural experiments;
3. Differences-in-differences-type strategies that use repeated observations to control for unobserved omitted factors.

The productive use of these basic techniques requires a solid conceptual foundation and a good understanding of the machinery of statistical inference. Both aspects of applied econometrics are covered here.

Our view of what's important has been shaped by our experience as empirical researchers, and especially by our work teaching and advising Economics Ph.D. students. This book was written with these students in mind. At the same time, we hope the book will find an audience among other groups of researchers who have an urgent need for practical answers regarding choice of technique and the interpretation of research findings. The concerns of applied econometrics are not fundamentally different from those in other social sciences or epidemiology. Anyone interested in using data to shape public policy or to promote public health must digest and use statistical results. Anyone interested in drawing useful inferences from data on people can be said to be an applied econometrician.

Many textbooks provide a guide to research methods and there is some overlap between this book and others in wide use. But our Companion differs from econometrics texts in a number of important ways. First, we believe that empirical research is most valuable when it uses data to answer specific causal questions, *as*

if in a randomized clinical trial. This view shapes our approach to all research questions. In the absence of a real experiment, we look for well-controlled comparisons and/or natural "quasi-experiments". Of course, some quasi-experimental research designs are more convincing than others, but the econometric methods used in these studies are almost always fairly simple. Consequently, our book is shorter and more focused than textbook treatments of econometric methods. We emphasize the conceptual issues and simple statistical techniques that turn up in the applied research we read and do, and illustrate these ideas and techniques with many empirical examples. Although our views of what's important are not universally shared among applied economists, there is no arguing with the fact that experimental and quasi-experimental research designs are increasingly at the heart of the most influential empirical studies in applied economics.

A second distinction we claim is a certain lack of seriousness. Most econometrics texts appear to take econometric models very seriously. Typically these books pay a lot of attention to the putative failures of classical modelling assumptions such as linearity and homoskedasticity. Warnings are sometimes issued. We take a more forgiving and less literal-minded approach. A principle that guides our discussion is that the estimators in common use almost always have a simple interpretation that is not heavily model-dependent. If the estimates you get are not the estimates you want, the fault lies in the econometrician and not the econometrics! A leading example is linear regression, which provides useful information about the conditional mean function regardless of the shape of this function. Likewise, instrumental variables methods estimate an average causal effect for a well-defined population even if the instrument does not affect everyone. The conceptual robustness of basic econometric tools is grasped intuitively by many applied researchers, but the theory behind this robustness does not feature in most texts. Our Companion also differs from most econometrics texts in that, on the inference side, we are not much concerned with asymptotic efficiency. Rather, our discussion of inference is devoted mostly to the finite-sample bugaboos that should bother practitioners.

The main prerequisites for the material here are basic training in probability and statistics. We especially hope that readers are comfortable with the elementary tools of statistical inference, such as t-statistics and standard errors. Familiarity with fundamental probability concepts like mathematical expectation is also helpful, but extraordinary mathematical sophistication is not required. Although important proofs are presented, the technical arguments are not very long or complicated. Unlike many upper-level econometrics texts, we go easy on the linear algebra. For this reason and others, our Companion should be an easier read than competing books. Finally, in the spirit of the Douglas Adams' lighthearted serial (*The Hitchhiker's Guide to the Galaxy* and *Mostly Harmless,* among others*)* from which we draw continued inspiration, our Companion may have occasional inaccuracies, but it is quite a bit cheaper than the many versions of the *Encyclopedia Galactica Econometrica* that dominate today's market. Grateful thanks to Princeton University Press for agreeing to distribute our Companion on these terms.

Acknowledgments

We had the benefit of comments from many friends and colleagues as this project progressed. Special thanks are due to Alberto Abadie, David Autor, Amitabh Chandra, Monica Chen, John DiNardo, Joe Doyle, Jerry Hausman, Andrea Ichino, Guido Imbens, Rafael Lalive, Alan Manning, Karen Norberg, Barbara Petrongolo, James Robinson, Tavneet Suri, and Jeff Wooldridge, who reacted to the draft at various stages. They are not to blame for our presumptuousness or remaining mistakes. Thanks also go to our students at LSE and MIT. They saw the material first and helped us decide what's important. We would especially like to acknowledge the skilled and tireless research assistance of Brigham Frandsen, Cynthia Kinnan, and Chris Smith. We're also grateful for the support and guidance of Tim Sullivan and Seth Ditchik, our editors at Princeton University Press. Last, certainly not least, we thank our wives for their love and support; they know better than anyone what it means to be an empiricist's companion.

Organization of this Book

We begin with two introductory chapters. The first describes the type of research agenda for which the material in subsequent chapters is most likely to be useful. The second discusses the sense in which experiments, i.e., randomized trials of the sort used in medical research, provide an ideal benchmark for the questions we find most interesting. After this introduction, the three chapters of Part II present core material on regression, instrumental variables, and differences-in-differences. These chapters emphasize both the universal properties of the relevant estimators (e.g., regression always approximates the conditional mean function) and the assumptions necessary for a causal interpretation of results (the conditional independence assumption; instruments as good as randomly assigned; parallel worlds). We then turn to important extensions in Part III. Chapter 6 covers regression discontinuity designs, which can be seen as either a variation on regression-control strategies or a type of instrumental variables strategy. In Chapter 7, we discuss the use of quantile regression for estimating effects on distributions. The last chapter covers important inference problems that are missed by the textbook asymptotic approach. Some chapters include more technical or specialized sections that can be skimmed or skipped without missing out on the main ideas - these are indicated with a star. Notation, an acronym glossary, and an index to empirical examples are gathered at the back of the book.

Part I

Introduction

Chapter 1

Questions about *Questions*

> 'I checked it very thoroughly,' said the computer, 'and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is.'
>
> Douglas Adams, *The Hitchhiker's Guide to the Galaxy* (1979)

Many econometrics courses are concerned with the details of empirical research, taking the choice of topic as given. But a coherent, interesting, and doable research agenda is the solid foundation on which useful statistical analyses are built. Good econometrics cannot save a shaky research agenda, but the promiscuous use of fancy econometric techniques sometimes brings down a good one. This chapter briefly discusses the basis for a successful research project. Like the biblical story of Exodus, a research agenda can be organized around four questions. We call these Frequently Asked Questions (FAQs), because they should be. The FAQs ask about the relationship of interest, the ideal experiment, the identification strategy, and the mode of inference.

In the beginning, we should ask: *What is the causal relationship of interest?* Although purely descriptive research has an important role to play, we believe that the most interesting research in social science is about cause and effect, like the effect of class size on children's test scores discussed in Chapters 2 and 6. A causal relationship is useful for making predictions about the consequences of changing circumstances or policies; it tells us what would happen in alternative (or "counterfactual") worlds. For example, as part of a research agenda investigating human productive capacity—what labor economists call human capital—we have both investigated the causal effect of schooling on wages (Card, 1999, surveys research in this area). The causal effect of schooling on wages is the increment to wages an individual would receive if he or she got more schooling. A range of studies suggest the causal effect of a college degree is about 40 percent higher wages on average, quite a payoff. The causal effect of schooling on wages is useful for predicting the earnings consequences of, say, changing the costs of attending college, or strengthening compulsory attendance laws. This relation is also of theoretical interest since it can be derived from an economic model.

As labor economists, we're most likely to study causal effects in samples of workers, but the unit of observation in causal research need not be an individual human being. Causal questions can be asked about firms, or, for that matter, countries. An example of the latter is Acemoglu, Johnson, and Robinson's (2001) research on the effect of colonial institutions on economic growth. This study is concerned with whether countries that inherited more democratic institutions from their colonial rulers later enjoyed higher economic growth as a consequence. The answer to this question has implications for our understanding of history and for the consequences of contemporary development policy. Today, for example, we might wonder whether newly forming democratic institutions are important for economic development in Iraq and Afghanistan. The case for democracy is far from clear-cut; at the moment, China is enjoying robust growth without the benefit of complete political freedom, while much of Latin America has democratized without a big growth payoff.

The second research FAQ is concerned with *the experiment that could ideally be used to capture the causal effect of interest.* In the case of schooling and wages, for example, we can imagine offering potential dropouts a reward for finishing school, and then studying the consequences. In fact, Angrist and Lavy (2007) have run just such an experiment. Although this study looks at short-term effects such as college enrollment, a longer-term follow-up might well look at wages. In the case of political institutions, we might like to go back in time and randomly assign different government structures to former colonies on their Independence Days (an experiment that is more likely to be made into a movie than to get funded by the National Science Foundation).

Ideal experiments are most often hypothetical. Still, hypothetical experiments are worth contemplating because they help us pick fruitful research topics. We'll support this claim by asking you to picture yourself as a researcher with no budget constraint and no Human Subjects Committee policing your inquiry for social correctness. Something like a well-funded Stanley Milgram, the psychologist who did path-breaking work on the response to authority in the 1960s using highly controversial experimental designs that would likely cost him his job today.

Seeking to understand the response to authority, Milgram (1963) showed he could convince experimental subjects to administer painful electric shocks to pitifully protesting victims (the shocks were fake and the victims were actors). This turned out to be controversial as well as clever—some psychologists claimed that the subjects who administered shocks were psychologically harmed by the experiment. Still, Milgram's study illustrates the point that there are many experiments we can think about, even if some are better left on the drawing board.[1] If you can't devise an experiment that answers your question in a world where anything goes, then the odds of generating useful results with a modest budget and non-experimental survey data seem pretty slim. The description of an ideal experiment also helps you formulate causal questions precisely.

[1] Milgram was later played by the actor William Shatner in a TV special, an honor that no economist has yet received, though Angrist is still hopeful.

The mechanics of an ideal experiment highlight the forces you'd like to manipulate and the factors you'd like to hold constant.

Research questions that cannot be answered by any experiment are FUQ'd: Fundamentally Unidentified Questions. What exactly does a FUQ'd question look like? At first blush, questions about the causal effect of race or gender seems like good candidates because these things are hard to manipulate in isolation ("imagine your chromosomes were switched at birth"). On the other hand, the issue economists care most about in the realm of race and sex, labor market discrimination, turns on whether someone treats you differently because they *believe* you to be black or white, male or female. The notion of a counterfactual world where men are perceived as women or vice versa has a long history and does not require Douglas-Adams-style outlandishness to entertain (Rosalind disguised as Ganymede fools everyone in Shakespeare's *As You Like It*). The idea of changing race is similarly near-fetched: In *The Human Stain*, Philip Roth imagines the world of Coleman Silk, a black Literature professor who passes as white in professional life. Labor economists imagine this sort of thing all the time. Sometimes we even construct such scenarios for the advancement of science, as in audit studies involving fake job applicants and resumes.[2]

A little imagination goes a long way when it comes to research design, but imagination cannot solve every problem. Suppose that we are interested in whether children do better in school by virtue of having started school a little older. Maybe the 7-year-old brain is better prepared for learning than the 6 year old brain. This question has a policy angle coming from the fact that, in an effort to boost test scores, some school districts are now entertaining older start-ages (to the chagrin of many working mothers). To assess the effects of delayed school entry on learning, we might randomly select some kids to start kindergarten at age 6, while others start at age 5, as is still typical. We are interested in whether those held back learn more in school, as evidenced by their elementary school test scores. To be concrete, say we look at test scores in first grade.

The problem with this question - the effects of start age on first grade test scores - is that the group that started school at age 7 is . . . older. And older kids tend to do better on tests, a pure maturation effect. Now, it might seem we can fix this by holding age constant instead of grade. Suppose we test those who started at age 6 in second grade and those who started at age 7 in first grade so everybody is tested at age 7. But the first group has spent more time in school; a fact that raises achievement if school is worth anything. There is no way to disentangle the start-age effect from maturation and time-in-school effects as long as kids are still in school. The problem here is that start age equals current age minus time in school. This deterministic link disappears in a sample of adults, so we might hope to investigate whether changes in entry-age policies affected adult outcomes like earnings or highest grade completed. But the effect of start age on elementary school test scores is most likely FUQ'd.

[2] A recent example is Bertrand and Mullainathan (2004) who compared employers' reponses to resumes with blacker-sounding and whiter-sounding first names, like Lakisha and Emily (though Fryer and Levitt, 2004, note that names may carry information about socioeconomic status as well as race.)

The third and fourth research FAQs are concerned with the nuts-and-bolts elements that produce a specific study. Question Number 3 asks: *what is your identification strategy?* Angrist and Krueger (1999) used the term *identification strategy* to describe the manner in which a researcher uses observational data (i.e., data not generated by a randomized trial) to approximate a real experiment. Again, returning to the schooling example, Angrist and Krueger (1991) used the interaction between compulsory attendance laws in American schools and students' season of birth as a natural experiment to estimate the effects of finishing high school on wages (season of birth affects the degree to which high school students are constrained by laws allowing them to drop out on their birthdays). Chapters 3-6 are primarily concerned with conceptual frameworks for identification strategies.

Although a focus on credible identification strategies is emblematic of modern empirical work, the juxtaposition of ideal and natural experiments has a long history in econometrics. Here is our econometrics forefather, Trygve Haavelmo (1944, p. 14)), appealing for more explicit discussion of both kinds of experimental designs:

> A design of experiments (a prescription of what the physicists call a "crucial experiment") is an essential appendix to any quantitative theory. And we usually have some such experiment in mind when we construct the theories, although—unfortunately—most economists do not describe their design of experiments explicitly. If they did, they would see that the experiments they have in mind may be grouped into two different classes, namely, (1) experiments that *we should like to make* to see if certain real economic phenomena—when artificially isolated from "other influences"—would verify certain hypotheses, and (2) the stream of experiments that Nature is steadily turning out from her own enormous laboratory, and which we merely watch as passive observers. In both cases the aim of the theory is the same, to become master of the happenings of real life.

The fourth research FAQ borrows language from Rubin (1991): *what is your mode of statistical inference?* The answer to this question describes the population to be studied, the sample to be used, and the assumptions made when constructing standard errors. Sometimes inference is straightforward, as when you use Census micro-data samples to study the American population. Often inference is more complex, however, especially with data that are clustered or grouped. The last chapter covers practical problems that arise once you've answered question number 4. Although inference issues are rarely very exciting, and often quite technical, the ultimate success of even a well-conceived and conceptually exciting project turns on the details of statistical inference. This sometimes-dispiriting fact inspired the following econometrics haiku, penned by then-econometrics-Ph.D.-student Keisuke Hirano on the occasion of completing his thesis:

> *T-stat looks too good.*
>
> *Use robust standard errors–*

significance gone.

As should be clear from the above discussion, the four research FAQs are part of a process of project development. The following chapters are concerned mostly with the econometric questions that come up after you've answered the research FAQs. In other words, issues that arise once your research agenda has been set. Before turning to the nuts and bolts of empirical work, however, we begin with a more detailed explanation of why randomized trials give us our benchmark.

Chapter 2

The Experimental Ideal

> It is an important and popular fact that things are not always what they seem. For instance, on the planet Earth, man had always assumed that he was more intelligent than dolphins because he had achieved so much—the wheel, New York, wars and so on—while all the dolphins had ever done was muck about in the water having a good time. But conversely, the dolphins had always believed that they were far more intelligent than man–for precisely the same reasons. In fact there was only one species on the planet more intelligent than dolphins, and they spent a lot of their time in behavioral research laboratories running round inside wheels and conducting frighteningly elegant and subtle experiments on man. The fact that once again man completely misinterpreted this relationship was entirely according to these creatures' plans.
>
> Douglas Adams, *The Hitchhiker's Guide to the Galaxy* (1979)

The most credible and influential research designs use random assignment. A case in point is the Perry preschool project, a 1962 randomized experiment designed to asses the effects of an early-intervention program involving 123 Black preschoolers in Ypsilanti (Michigan). The Perry treatment group was randomly assigned to an intensive intervention that included preschool education and home visits. It's hard to exaggerate the impact of the small but well-designed Perry experiment, which generated follow-up data through 1993 on the participants at age 27. Dozens of academic studies cite or use the Perry findings (see, e.g., Barnett, 1992). Most importantly, the Perry project provided the intellectual basis for the massive Head Start pre-school program, begun in 1964, which ultimately served (and continues to serve) millions of American children.[1]

[1] The Perry data continue to get attention, particular as policy-interest has returned to early education. A recent re-analysis by Michael Anderson (2006) confirms many of the findings from the original Perry study, though Anderson also shows that the overall positive effects of Perry are driven entirely by the impact on girls. The Perry intervention seems to have done nothing for boys.

2.1 The Selection Problem

We take a brief time-out for a more formal discussion of the role experiments play in uncovering causal effects. Suppose you are interested in a causal "if-then" question. To be concrete, consider a simple example: Do hospitals make people healthier? For our purposes, this question is allegorical, but it is surprisingly close to the sort of causal question health economists care about. To make this question more realistic, imagine we're studying a poor elderly population that uses hospital emergency rooms for primary care. Some of these patients are admitted to the hospital. This sort of care is expensive, crowds hospital facilities, and is, perhaps, not very effective (see, e.g., Grumbach, Keane, and Bindman, 1993). In fact, exposure to other sick patients by those who are themselves vulnerable might have a net negative impact on their health.

Since those admitted to the hospital get many valuable services, the answer to the hospital-effectiveness question still seems likely to be "yes". But will the data back this up? The natural approach for an empirically-minded person is to compare the health status of those who have been to the hospital to the health of those who have not. The National Health Interview Survey (NHIS) contains the information needed to make this comparison. Specifically, it includes a question "During the past 12 months, was the respondent a patient in a hospital overnight?" which we can use to identify recent hospital visitors. The NHIS also asks "Would you say your health in general is excellent, very good, good, fair, poor?" The following table displays the mean health status (assigning a 1 to excellent health and a 5 to poor health) among those who have been hospitalized and those who have not (tabulated from the 2005 NHIS):

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003

The difference in the means is 0.71, a large and highly significant contrast in favor of the *non-hospitalized*, with a t-statistic of 58.9.

Taken at face value, this result suggests that going to the hospital makes people sicker. It's not impossible this is the right answer: hospitals are full of other sick people who might infect us, and dangerous machines and chemicals that might hurt us. Still, it's easy to see why this comparison should not be taken at face value: people who go to the hospital are probably less healthy to begin with. Moreover, even after hospitalization people who have sought medical care are not as healthy, on average, as those who never get hospitalized in the first place, though they may well be better than they otherwise would have been.

To describe this problem more precisely, think about hospital treatment as described by a binary random variable, $\text{D}_i = \{0, 1\}$. The outcome of interest, a measure of health status, is denoted by Y_i. The question is whether Y_i is *affected* by hospital care. To address this question, we assume we can imagine what might have happened to someone who went to the hospital if they had not gone and vice versa. Hence, for any individual there are two potential health variables:

$$\textit{potential outcome} = \begin{cases} \mathrm{Y}_{1i} & \text{if } \mathrm{D}_i = 1 \\ \mathrm{Y}_{0i} & \text{if } \mathrm{D}_i = 0 \end{cases}.$$

In other words, Y_{0i} is the health status of an individual had he not gone to the hospital, irrespective of whether he actually went, while Y_{1i} is the individual's health status if he goes. We would like to know the difference between Y_{1i} and Y_{0i}, which can be said to be the causal effect of going to the hospital for individual i. This is what we would measure if we could go back in time and change a person's treatment status.[2]

The observed outcome, Y_i, can be written in terms of potential outcomes as

$$\begin{aligned} \mathrm{Y}_i &= \begin{cases} \mathrm{Y}_{1i} & \text{if } \mathrm{D}_i = 1 \\ \mathrm{Y}_{0i} & \text{if } \mathrm{D}_i = 0 \end{cases} \\ &= \mathrm{Y}_{0i} + (\mathrm{Y}_{1i} - \mathrm{Y}_{0i})\mathrm{D}_i. \end{aligned} \tag{2.1.1}$$

This notation is useful because $\mathrm{Y}_{1i} - \mathrm{Y}_{0i}$ is the causal effect of hospitalization for an individual. In general, there is likely to be a distribution of both Y_{1i} and Y_{0i} in the population, so the treatment effect can be different for different people. But because we never see both potential outcomes for any one person, we must learn about the effects of hospitalization by comparing the average health of those who were and were not hospitalized.

A naive comparison of averages by hospitalization status tells us something about potential outcomes, though not necessarily what we want to know. The comparison of average health conditional on hospitalization status is formally linked to the average causal effect by the equation below:

$$\underbrace{E\left[\mathrm{Y}_i|\mathrm{D}_i = 1\right] - E[\mathrm{Y}_i|\mathrm{D}_i = 0]}_{\text{Observed difference in average health}} = \underbrace{E\left[\mathrm{Y}_{1i}|\mathrm{D}_i = 1\right] - E[\mathrm{Y}_{0i}|\mathrm{D}_i = 1]}_{\text{average treatment effect on the treated}} + \underbrace{E\left[\mathrm{Y}_{0i}|\mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_{0i}|\mathrm{D}_i = 0\right]}_{\text{selection bias}}$$

The term

$$E[\mathrm{Y}_{1i}|\mathrm{D}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{D}_i = 1] = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1]$$

is the *average causal effect of hospitalization on those who were hospitalized*. This term captures the averages difference between the health of the hospitalized, $E[\mathrm{Y}_{1i}|\mathrm{D}_i = 1]$, and what would have happened to *them* had they not been hospitalized, $E[\mathrm{Y}_{0i}|\mathrm{D}_i = 1]$. The observed difference in health status however, adds to this causal effect a term called *selection bias*. This term is the difference in average Y_{0i} between those who

[2] The potential outcomes idea is a fundamental building block in modern research on causal effects. Important references developing this idea are Rubin (1974, 1977), and Holland (1986), who refers to a causal framework involving potential outcomes as the Rubin Causal Model.

were and were not hospitalized. Because the sick are more likely than the healthy to seek treatment, those who were hospitalized have worse Y_{0i}'s, making selection bias negative in this example. The selection bias may be so large (in absolute value) that it completely masks a positive treatment effect. The goal of most empirical economic research is to overcome selection bias, and therefore to say something about the causal effect of a variable like D_i.

2.2 Random Assignment Solves the Selection Problem

Random assignment of D_i solves the selection problem because random assignment makes D_i independent of potential outcomes. To see this, note that

$$\begin{aligned} E[\text{Y}_i|\text{D}_i = 1] - E[\text{Y}_i|\text{D}_i = 0] &= E[\text{Y}_{1i}|\text{D}_i = 1] - E[\text{Y}_{0i}|\text{D}_i = 0] \\ &= E[\text{Y}_{1i}|\text{D}_i = 1] - E[\text{Y}_{0i}|\text{D}_i = 1], \end{aligned}$$

where the independence of Y_{0i} and D_i allows us to swap $E[\text{Y}_{0i}|\text{D}_i = 1]$ for $E[\text{Y}_{0i}|\text{D}_i = 0]$ in the second line. In fact, given random assignment, this simplifies further to

$$\begin{aligned} E\left[\text{Y}_{1i}|\text{D}_i = 1\right] - E\left[\text{Y}_{0i}|\text{D}_i = 1\right] &= E\left[\text{Y}_{1i} - \text{Y}_{0i}|\text{D}_i = 1\right] \\ &= E\left[\text{Y}_{1i} - \text{Y}_{0i}\right]. \end{aligned}$$

The effect of randomly-assigned hospitalization on the hospitalized is the same as the effect of hospitalization on a randomly chosen patient. The main thing, however, is that random assignment of D_i eliminates selection bias. This does not mean that randomized trials are problem-free, but in principle they solve the most important problem that arises in empirical research.

How relevant is our hospitalization allegory? Experiments often reveal things that are not what they seem on the basis of naive comparisons alone. A recent example from medicine is the evaluation of hormone replacement therapy (HRT). This is a medical intervention that was recommended for middle-aged women to reduce menopausal symptoms. Evidence from the Nurses Health Study, a large and influential non-experimental survey of nurses, showed better health among the HRT users. In contrast, the results of a recently completed randomized trial shows few benefits of HRT. What's worse, the randomized trial revealed serious side effects that were not apparent in the non-experimental data (see, e.g., Women's Health Initiative [WHI], Hsia, *et al.*, 2006).

An iconic example from our own field of labor economics is the evaluation of government-subsidized training programs. These are programs that provide a combination of classroom instruction and on-the-job training for groups of disadvantaged workers such as the long-term unemployed, drug addicts, and ex-offenders. The idea is to increase employment and earnings. Paradoxically, studies based on non-

experimental comparisons of participants and non-participants often show that after training, the trainees earn less than plausible comparison groups (see, e.g., Ashenfelter, 1978; Ashenfelter and Card, 1985; Lalonde 1995). Here too, selection bias is a natural concern since subsidized training programs are meant to serve men and women with low earnings potential. Not surprisingly, therefore, simple comparisons of program participants with non-participants often show lower earnings for the participants. In contrast, evidence from randomized evaluations of training programs generate mostly positive effects (see, e.g., Lalonde, 1986; Orr, et al, 1996).

Randomized trials are not yet as common in social science as in medicine but they are becoming more prevalent. One area where the importance of random assignment is growing rapidly is education research (Angrist, 2004). The 2002 Education Sciences Reform Act passed by the U.S. Congress mandates the use of rigorous experimental or quasi-experimental research designs for all federally-funded education studies. We can therefore expect to see many more randomized trials in education research in the years to come. A pioneering randomized study from the field of education is the Tennessee STAR experiment designed to estimate the effects of smaller classes in primary school.

Labor economists and others have a long tradition of trying to establish causal links between features of the classroom environment and children's learning, an area of investigation that we call "education production." This terminology reflects the fact that we think of features of the school environment as inputs that cost money, while the output that schools produce is student learning. A key question in research on education production is which inputs produce the most learning given their costs. One of the most expensive inputs is class size - since smaller classes can only be had by hiring more teachers. It is therefore important to know whether the expense of smaller classes has a payoff in terms of higher student achievement. The STAR experiment was meant to answer this question.

Many studies of education production using non-experimental data suggest there is little or no link between class size and student learning. So perhaps school systems can save money by hiring fewer teachers with no consequent reduction in achievement. The observed relation between class size and student achievement should not be taken at face value, however, since weaker students are often deliberately grouped into smaller classes. A randomized trial overcomes this problem by ensuring that we are comparing apples to apples, i.e., that the students assigned to classes of different sizes are otherwise comparable. Results from the Tennessee STAR experiment point to a strong and lasting payoff to smaller classes (see Finn and Achilles, 1990, for the original study, and Krueger, 1999, for an econometric analysis of the STAR data).

The STAR experiment was unusually ambitious and influential, and therefore worth describing in some detail. It cost about $12 million and was implemented for a cohort of kindergartners in 1985/86. The study ran for four years, i.e. until the original cohort of kindergartners was in third grade, and involved about 11,600 children. The average class size in regular Tennessee classes in 1985/86 was about 22.3. The experiment assigned students to one of three treatments: small classes with 13-17 children, regular classes

with 22-25 children and a part-time teacher's aide, or regular classes with a full time teacher's aide. Schools with at least three classes in each grade could choose to participate in the experiment.

The first question to ask about a randomized experiment is whether the randomization successfully balanced subject's characteristics across the different treatment groups. To assess this, it's common to compare pre-treatment outcomes or other covariates across groups. Unfortunately, the STAR data fail to include any pre-treatment test scores, though it is possible to look at characteristics of children such as race and age. Table 2.2.1, reproduced from Krueger (1999), compares the means of these variables. The student

Table 2.2.1: Comparison of treatment and control characteristics in the Tennessee STAR experiment

	Students who entered STAR in kindergarten				
	Variable	Small	Regular	Regular/Aide	Joint P-value
1.	Free lunch	.47	.48	.50	.09
2.	White/Asian	.68	.67	.66	.26
3.	Age in 1985	5.44	5.43	5.42	.32
4.	Attrition rate	.49	.52	.53	.02
5.	Class size in kindergarten	15.10	22.40	22.80	.00
6.	Percentile score in kindergarten	54.70	48.90	50.00	.00

Notes: Adapted from Krueger (1999), Table 1. The table shows means of variables by treatment status. The P-value in the last column is for the F-test of equality of variable means across all three groups. All variables except attrition are for the first year a student is observed, The free lunch variable is the fraction receiving a free lunch. The percentile score is the average percentile score on three Stanford Achievement Tests. The attrition rate is the proportion lost to follow up before completing third grade.

characteristics in the table are a free lunch variable, student race, and student age. Free lunch status is a good measure of family income, since only poor children qualify for a free school lunch. Differences in these characteristics across the three class types are small and none are significantly different from zero. This suggests the random assignment worked as intended.

Table 2.2.1 also presents information on average class size, the attrition rate, and test scores, measured here on a percentile scale. The attrition rate was lower in small kindergarten classrooms. This is potential a problem, at least in principle.[3] Class sizes are significantly lower in the assigned-to-be-small class rooms, which means that the experiment succeeded in creating the desired variation. If many of the parents of children assigned to regular classes had effectively lobbied teachers and principals to get their children assigned to small classes, the gap in class size across groups would be much smaller.

[3]Krueger (1999) devotes considerable attention to the attrition problem. Differences in attrition rates across groups may result in a sample of students in higher grades that is not randomly distributed across class types. The kindergarten results, which were unaffected by attrition, are therefore the most reliable.

Because randomization eliminates selection bias, the difference in outcomes across treatment groups captures the average causal effect of class size (relative to regular classes with a part-time aide). In practice, the difference in means between treatment and control groups can be obtained from a regression of test scores on dummies for each treatment group, a point we expand on below. The estimated treatment-control differences for kindergartners, reported in Table 2.2.2 (derived from Krueger, 1999, Table 5), show a small-class effect of about 5 to 6 percentile points. The effect size is about $.2\sigma$, where σ is the standard deviation of the percentile score in kindergarten. The small-class effect is significantly different from zero, while the

Table 2.2.2: Experimental estimates of the effect of class-size assignment on test scores

Explanatory variable	(1)	(2)	(3)	(4)
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes)	–	–	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	–	–	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	–	–	-13.15	-13.07
			(.77)	(.77)
White teacher	–	–	–	-.57
				(2.10)
Teacher experience	–	–	–	.26
				(.10)
Master's degree	–	–	–	-0.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Note: Adapted from Krueger (1999), Table 5. The dependent variable is the Stanford Achievement Test percentile score. Robust standard errors that allow for correlated residuals within classes are shown in parentheses. The sample size is 5681.

regular/aide effect is small and insignificant.

The STAR study, an exemplary randomized trial in the annals of social science, also highlights the logistical difficulty, long duration, and potentially high cost of randomized trials. In many cases, such trials are impractical.[4] In other cases, we would like an answer sooner rather than later. Much of the research

[4] Randomized trials are never perfect and STAR is no exception. Pupils who repeated or skipped a grade left the experiment. Students who entered an experimental school one grade later were added to the experiment and randomly assigned to one of the classes. One unfortunate aspect of the experiment is that students in the regular and regular/aide classes were reassigned after the kindergarten year, possibly due to protests of the parents with children in the regular classrooms. There was also some switching of children after the kindergarten year. Despite these problems, the STAR experiment seems to have been an

we do, therefore, attempts to exploit cheaper and more readily available sources of variation. We hope to find natural or quasi-experiments that mimic a randomized trial by changing the variable of interest while other factors are kept balanced. Can we always find a convincing natural experiment? Of course not. Nevertheless, we take the position that a notional randomized trial is our benchmark. Not all researchers share this view, but many do. We heard it first from our teacher and thesis advisor, Orley Ashenfelter, a pioneering proponent of experiments and quasi-experimental research designs in social science. Here is Ashenfelter (1991) assessing the credibility of the observational studies linking schooling and income:

> How convincing is the evidence linking education and income? Here is my answer: Pretty convincing. If I had to bet on what an ideal experiment would indicate, I bet that it would show that better educated workers earn more.

The quasi-experimental study of class size by Angrist and Lavy (1999) illustrates the manner in which non-experimental data can be analyzed in an experimental spirit. The Angrist and Lavy study relies on the fact that in Israel, class size is capped at 40. Therefore, a child in a fifth grade cohort of 40 students ends up in a class of 40 while a child in fifth grade cohort of 41 students ends up in a class only half as large because the cohort is split. Since students in cohorts of size 40 and 41 are likely to be similar on other dimensions such as ability and family background, we can think of the difference between 40 and 41 students enrolled as being "as good as randomly assigned."

The Angrist-Lavy study compares students in grades with enrollments above and below the class-size cutoffs to construct well-controlled estimates of the effects of a sharp change in class size without the benefit of a real experiment. As in Tennessee STAR, the Angrist and Lavy (1999) results point to a strong link between class size and achievement. This is in marked contrast with naive analyses, also reported by Angrist and Lavy, based on simple comparisons between those enrolled in larger and smaller classes. These comparisons show students in smaller classes doing worse on standardized tests. The hospital allegory of selection bias would therefore seem to apply to the class-size question as well.[5]

2.3 Regression Analysis of Experiments

Regression is a useful tool for the study of causal questions, including the analysis of data from experiments. Suppose (for now) that the treatment effect is the same for everyone, say $\mathrm{Y}_{1i} - \mathrm{Y}_{0i} = \rho$, a constant. With

extremely well implemented randomized trial. Krueger's (1999) analysis suggests that none of these implementation problems affected the main conclusions of the study.

[5] The Angrist-Lavy (1999) results turn up again in Chapter 6, as an illustration of the quasi-experimental regression-discontinuity research design.

constant treatment effects, we can rewrite equation (2.1.1) in the form

$$\begin{array}{ccccccc} \mathrm{Y}_i = & \alpha & + & \rho & \mathrm{D}_i & + & \eta_i, \\ & \shortparallel & & \shortparallel & & & \shortparallel \\ & E(\mathrm{Y}_{0i}) & & (\mathrm{Y}_{1i} - \mathrm{Y}_{0i}) & & & \mathrm{Y}_{0i} - E(\mathrm{Y}_{0i}) \end{array} \tag{2.3.1}$$

where η_i is the random part of Y_{0i}. Evaluating the conditional expectation of this equation with treatment status switched off and on gives

$$\begin{aligned} E\left[\mathrm{Y}_i|\mathrm{D}_i = 1\right] &= \alpha + \rho + E\left[\eta_i|\mathrm{D}_i = 1\right] \\ E\left[\mathrm{Y}_i|\mathrm{D}_i = 0\right] &= \alpha + E\left[\eta_i|\mathrm{D}_i = 0\right], \end{aligned}$$

so that,

$$E[\mathrm{Y}_i|\mathrm{D}_i = 1] - E[\mathrm{Y}_i|\mathrm{D}_i = 0] = \underbrace{\rho}_{\text{treatment effect}} + \underbrace{E[\eta_i|\mathrm{D}_i = 1] - E[\eta_i|\mathrm{D}_i = 0]}_{\text{selection bias}}.$$

Thus, selection bias amounts to correlation between the regression error term, η_i, and the regressor, D_i. Since

$$E\left[\eta_i|\mathrm{D}_i = 1\right] - E\left[\eta_i|\mathrm{D}_i = 0\right] = E\left[\mathrm{Y}_{0i}|\mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_{0i}|\mathrm{D}_i = 0\right],$$

this correlation reflects the difference in (no-treatment) potential outcomes between those who get treated and those who don't. In the hospital allegory, those who were treated had poorer health outcomes in the no-treatment state, while in the Angrist and Lavy (1999) study, students in smaller classes tend to have intrinsically lower test scores.

In the STAR experiment, where D_i is randomly assigned, the selection term disappears, and a regression of Y_i on D_i estimates the causal effect of interest, ρ. The remainder of Table 2.2.2 shows different regression specifications, some of which include covariates other than the random assignment indicator, D_i. Covariates play two roles in regression analyses of experimental data. First, the STAR experimental design used conditional random assignment. In particular, assignment to classes of different sizes was random within schools, but not across schools. Students attending schools of different types (say, urban versus rural) were a bit more or less likely to be assigned to a small class. The comparison in column 1 of Table 2.2.2, which makes no adjustment for this, might therefore be contaminated by differences in achievement in schools of different types. To adjust for this, some of Krueger's regression models include school fixed effects, i.e., a separate intercept for each school in the STAR data. In practice, the consequences of adjusting for school

fixed effects is rather minor, but we wouldn't know this without taking a look. We will have more to say about regression models with fixed effects in Chapter 5.

The other controls in Krueger's table describe student characteristics such as race, age, and free lunch status. We saw before that these individual characteristics are balanced across class types, i.e. they are not systematically related to the class-size assignment of the student. If these controls, call them X_i, are uncorrelated with the treatment D_i, then they will not affect the estimate of ρ. In other words, estimates of ρ in the long regression,

$$\mathrm{Y}_i = \alpha + \rho \mathrm{D}_i + \mathrm{X}_i'\gamma + \eta_i \tag{2.3.2}$$

will be close to estimates of ρ in the short regression, (2.3.1). This is a point we expand on in Chapter 3.

Nevertheless, inclusion of the variables X_i may generate more precise estimates of the causal effect of interest. Notice that the standard error of the estimated treatment effects in column 3 is smaller than the corresponding standard error in column 2. Although the control variables, X_i, are uncorrelated with D_i, they have substantial explanatory power for Y_i. Including these control variables therefore reduces the residual variance, which in turn lowers the standard error of the regression estimates. Similarly, the standard errors of the estimates of ρ are reduced by the inclusion of school fixed effects because these too explain an important part of the variance in student performance. The last column adds teacher characteristics. Because teachers were randomly assigned to classes, and teacher characteristics appear to have little to do with student achievement in these data, both the estimated effect of small classes and it's standard error are unchanged by the addition of teacher variables.

Regression plays an exceptionally important role in empirical economic research. Some regressions are simply descriptive tools, as in much of the research on earnings inequality. As we've seen in this chapter, regression is well-suited to the analysis of experimental data. In some cases, regression can also be used to approximate experiments in the absence of random assignment. But before we can get into the important question of when a regression is likely to have a causal interpretation, it is useful to review a number of fundamental regression facts and properties. These facts and properties are reliably true for any regression, regardless of your purpose in running it.

Part II

The Core

Chapter 3

Making Regression Make Sense

> 'Let us think the unthinkable, let us do the undoable.
> Let us prepare to grapple with the ineffable itself,
> and see if we may not eff it after all.'
> Douglas Adams, *Dirk Gently's Holistic Detective Agency* (1990)

Angrist recounts:

> I ran my first regression in the summer of 1979 between my freshman and sophomore years as a student at Oberlin College. I was working as a research assistant for Allan Meltzer and Scott Richard, faculty members at Carnegie-Mellon University, near my house in Pittsburgh. I was still mostly interested in a career in special education, and had planned to go back to work as an orderly in a state mental hospital, my previous summer job. But Econ 101 had got me thinking, and I could also see that at the same wage rate, a research assistant's hours and working conditions were better than those of a hospital orderly. My research assistant duties included data collection and regression analysis, though I did not understand regression or even statistics at the time.
>
> The paper I was working on that summer (Meltzer and Richard, 1983), is an attempt to link the size of governments in democracies, measured as government expenditure over GDP, to income inequality. Most income distributions have a long right tail, which means that average income tends to be way above the median. When inequality grows, more voters find themselves with below-average incomes. Annoyed by this, those with incomes between the median and the average may join those with incomes below the median in voting for fiscal policies which - following Robin Hood - take from the rich and give to the poor. The size of government consequently increases.
>
> I absorbed the basic theory behind the Meltzer and Richards project, though I didn't find it

all that plausible, since voter turnout is low for the poor. I also remember arguing with Alan Meltzer over whether government expenditure on education should be classified as a public good (something that benefits everyone in society as well as those directly affected) or a private good publicly supplied, and therefore a form of redistribution like welfare. You might say this project marked the beginning of my interest in the social returns to education, a topic I went back to with more enthusiasm and understanding in Acemoglu and Angrist (2000).

Today, I understand the Meltzer and Richard (1983) study as an attempt to use regression to uncover and quantify an interesting causal relation. At the time, however, I was purely a regression mechanic. Sometimes I found the RA work depressing. Days would go by where I didn't talk to anybody but my bosses and the occasional Carnegie-Mellon Ph.D. student, most of whom spoke little English anyway. The best part of the job was lunch with Alan Meltzer, a distinguished scholar and a patient and good-natured supervisor, who was happy to chat while we ate the contents of our brown-bags (this did not take long as Allan ate little and I ate fast). I remember asking Allan whether he found it satisfying to spend his days perusing regression output, which then came on reams of double-wide green-bar paper. Meltzer laughed and said there was nothing he would rather be doing.

Now, we too spend our days (at least, the good ones) happily perusing regression output, in the manner of our teachers and advisors in college and graduate school. This chapter explains why.

3.1 Regression Fundamentals

The end of the previous chapter introduces regression models as a computational device for the estimation of treatment-control differences in an experiment, with and without covariates. Because the regressor of interest in the class size study discussed in Section 2.3 was randomly assigned, the resulting estimates have a causal interpretation. In most cases, however, regression is used with observational data. Without the benefit of random assignment, regression estimates may or may not have a causal interpretation. We return to the central question of what makes a regression causal later in this chapter.

Setting aside the relatively abstract causality problem for the moment, we start with the mechanical properties of regression estimates. These are universal features of the population regression vector and its sample analog that have nothing to do with a researcher's interpretation of his output. This chapter begins by reviewing these properties, which include:

(i) the intimate connection between the population regression function and the conditional expectation function

(ii) how and why regression coefficients change as covariates are added or removed from the model

(iii) the close link between regression and other "control strategies" such as matching

(iv) the sampling distribution of regression estimates

3.1.1 Economic Relationships and the Conditional Expectation Function

Empirical economic research in our field of Labor Economics is typically concerned with the statistical analysis of individual economic circumstances, and especially differences between people that might account for differences in their economic fortunes. Such differences in economic fortune are notoriously hard to explain; they are, in a word, random. As applied econometricians, however, we believe we can summarize and interpret randomness in a useful way. An example of "systematic randomness" mentioned in the introduction is the connection between education and earnings. On average, people with more schooling earn more than people with less schooling. The connection between schooling and average earnings has considerable predictive power, in spite of the enormous variation in individual circumstances that sometimes clouds this fact. Of course, the fact that more educated people earn more than less educated people does not mean that schooling *causes* earnings to increase. The question of whether the earnings-schooling relationship is causal is of enormous importance, and we will come back to it many times. Even without resolving the difficult question of causality, however, it's clear that education predicts earnings in a narrow statistical sense. This predictive power is compellingly summarized by the conditional expectation function (CEF).

The CEF for a dependent variable, Y_i given a $\mathrm{K}\times 1$ vector of covariates, X_i (with elements x_{ki}) is the expectation, or population average of Y_i with X_i held fixed. The population average can be thought of as the mean in an infinitely large sample, or the average in a completely enumerated finite population. The CEF is written $E\left[\mathrm{Y}_i|\mathrm{X}_i\right]$ and is a function of X_i. Because X_i is random, the CEF is random, though sometimes we work with a particular value of the CEF, say $E[\mathrm{Y}_i|\mathrm{X}_i{=}42]$, assuming 42 is a possible value for X_i. In Chapter 2, we briefly considered the CEF $E[\mathrm{Y}_i|\mathrm{D}_i]$, where D_i is a zero-one variable. This CEF takes on two values, $E[\mathrm{Y}_i|\mathrm{D}_i = 1]$ and $E[\mathrm{Y}_i|\mathrm{D}_i = 0]$. Although this special case is important, we are most often interested in CEFs that are functions of many variables, conveniently subsumed in the vector, X_i. For a specific value of X_i, say $\mathrm{X}_i = x$, we write $E\left[\mathrm{Y}_i|\mathrm{X}_i = x\right]$. For continuous Y_i with conditional density $f_y\left(\cdot|\mathrm{X}_i = x\right)$, the CEF is

$$E\left[\mathrm{Y}_i|\mathrm{X}_i = x\right] = \int t f_y\left(t|\mathrm{X}_i = x\right) dt.$$

If Y_i is discrete, $E\left[\mathrm{Y}_i|\mathrm{X}_i = x\right]$ equals the sum $\sum_t t f_y\left(t|\mathrm{X}_i = x\right)$.

Expectation is a population concept. In practice, data usually come in the form of samples and rarely consist of an entire population. We therefore use samples to make inferences about the population. For example, the sample CEF is used to learn about the population CEF. This is always necessary but we postpone a discussion of the formal inference step taking us from sample to population until Section 3.1.3. Our "population first" approach to econometrics is motivated by the fact that we must define the objects of

interest before we can use data to study them.[1]

Figure 3.1.1 plots the CEF of log weekly wages given schooling for a sample of middle-aged white men from the 1980 Census. The distribution of earnings is also plotted for a few key values: 4, 8, 12, and 16 years of schooling. The CEF in the figure captures the fact that—the enormous variation individual circumstances notwithstanding—people with more schooling generally earn more, on average. The average earnings gain associated with a year of schooling is typically about 10 percent.

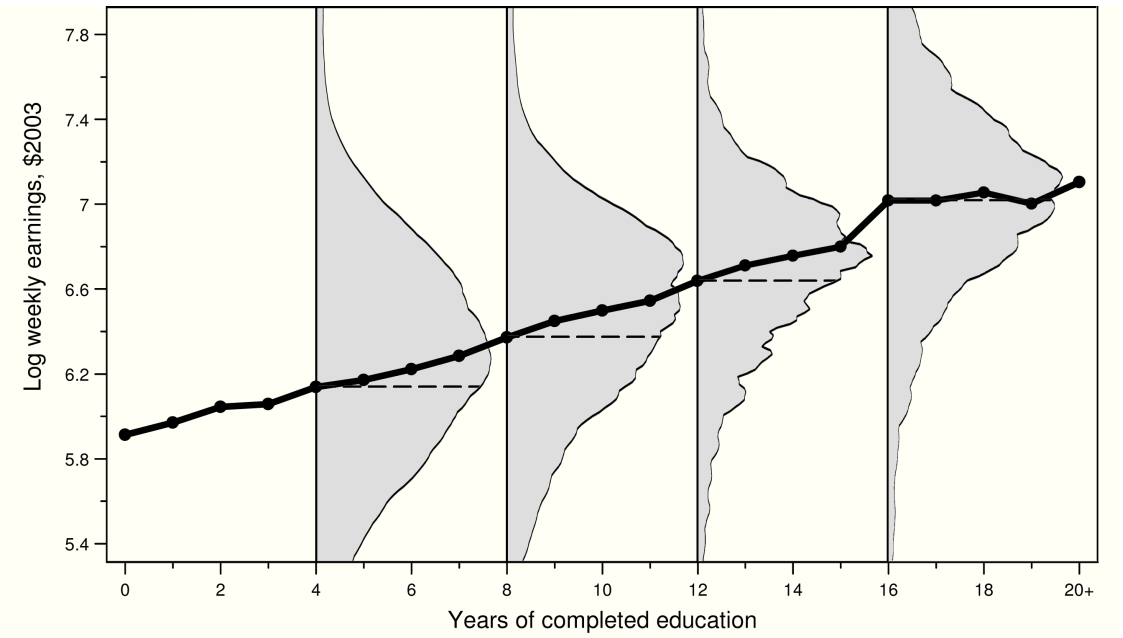

Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49 in the 1980 IPUMS 5 percent file.

An important complement to the CEF is the law of iterated expectations. This law says that an unconditional expectation can be written as the population average of the CEF. In other words

$$E\left[\mathrm{Y}_i\right] = E\{E\left[\mathrm{Y}_i|\mathrm{X}_i\right]\}, \tag{3.1.1}$$

where the outer expectation uses the distribution of X_i. Here is proof of the law of iterated expectations for continuously distributed $(\mathrm{X}_i, \mathrm{Y}_i)$ with joint density $f_{xy}(u,t)$, where $f_y(t|\mathrm{X}_i = x)$ is the conditional

[1] Examples of pedagogical writing using the "population-first" approach to econometrics include Chamberlain (1984), Goldberger (1991), and Manski (1991).

distribution of Y_i given $\mathrm{X}_i = x$ and $g_y(t)$ and $g_x(u)$ are the marginal densities:

$$\begin{aligned} E\{E\left[\mathrm{Y}_i|\mathrm{X}_i\right]\} &= \int E\left[\mathrm{Y}_i|\mathrm{X}_i = u\right] g_x(u)du \\ &= \int \left[\int t f_y\left(t|\mathrm{X}_i = u\right) dt\right] g_x(u)du \\ &= \int\int t f_y\left(t|\mathrm{X}_i = u\right) g_x(u)dudt \\ &= \int t\left[\int f_y\left(t|\mathrm{X}_i = u\right) g_x(u)du\right] dt = \int t\left[\int f_{xy}\left(u,t\right) du\right] dt \\ &= \int t g_y(t)dt. \end{aligned}$$

The integrals in this derivation run over the possible values of X_i and Y_i (indexed by u and t). We've laid out these steps because the CEF and its properties are central to the rest of this chapter.

The power of the law of iterated expectations comes from the way it breaks a random variable into two pieces.

Theorem 3.1.1 *The CEF-Decomposition Property*

$$\mathrm{Y}_i = E\left[\mathrm{Y}_i|\mathrm{X}_i\right] + \varepsilon_i,$$

where (i) ε_i is mean-independent of X_i, i.e., $E[\varepsilon_i|\mathrm{X}_i] = 0$,and, therefore, (ii) ε_i is uncorrelated with any function of X_i.

Proof. (i) $E[\varepsilon_i|\mathrm{X}_i] = E[\mathrm{Y}_i - E\left[\mathrm{Y}_i|\mathrm{X}_i\right]|\ \mathrm{X}_i] = E\left[\mathrm{Y}_i|\mathrm{X}_i\right] - E\left[\mathrm{Y}_i|\mathrm{X}_i\right] = 0$;(ii) This follows from (i): Let $h(\mathrm{X}_i)$ be any function of X_i. By the law of iterated expectations, $E[h(\mathrm{X}_i)\varepsilon_i] = E\{h(\mathrm{X}_i)E[\varepsilon_i|\mathrm{X}_i]\}$ and by mean-independence, $E[\varepsilon_i|\mathrm{X}_i] = 0$. ■

This theorem says that any random variable, Y_i, can be decomposed into a piece that's "explained by X_i", i.e., the CEF, and a piece left over which is orthogonal to (i.e., uncorrelated with) any function of X_i.

The CEF is a good summary of the relationship between Y_i and X_i for a number of reasons. First, we are used to thinking of averages as providing a representative value for a random variable. More formally, the CEF is the best predictor of Y_i given X_i in the sense that it solves a Minimum Mean Squared Error (MMSE) prediction problem. This CEF-prediction property is a consequence of the CEF-decomposition property:

Theorem 3.1.2 *The CEF-Prediction Property.*

Let $m\left(\mathrm{X}_i\right)$ be any function of X_i. The CEF solves

$$E\left[\mathrm{Y}_i|\mathrm{X}_i\right] = \underset{m(\mathrm{X}_i)}{\arg\min} E\left[\left(\mathrm{Y}_i - m\left(\mathrm{X}_i\right)\right)^2\right],$$

so it is the MMSE predictor of Y_i given X_i.

Proof. Write

$$\begin{aligned}(\mathrm{Y}_i - m(\mathrm{X}_i))^2 &= ((\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]) + (E[\mathrm{Y}_i|\mathrm{X}_i] - m(\mathrm{X}_i)))^2 \\ &= (\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i])^2 + 2(E[\mathrm{Y}_i|\mathrm{X}_i] - m(\mathrm{X}_i))(\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]) \\ &\quad + (E[\mathrm{Y}_i|\mathrm{X}_i] - m(\mathrm{X}_i))^2\end{aligned}$$

The first term doesn't matter because it doesn't involve $m(\mathrm{X}_i)$. The second term can be written $h(\mathrm{X}_i)\varepsilon_i$, where $h(\mathrm{X}_i) \equiv 2(E[\mathrm{Y}_i|\mathrm{X}_i] - m(\mathrm{X}_i))$, and therefore has expectation zero by the CEF-decomposition property. The last term is minimized at zero when $m(\mathrm{X}_i)$ is the CEF. ■

A final property of the CEF, closely related to both the CEF decomposition and prediction properties, is the Analysis-of-Variance (ANOVA) Theorem:

Theorem 3.1.3 *The ANOVA Theorem*

$$V(\mathrm{Y}_i) = V(E[\mathrm{Y}_i|\mathrm{X}_i]) + E[V(\mathrm{Y}_i|\mathrm{X}_i)]$$

where $V(\cdot)$ denotes variance and $V(\mathrm{Y}_i|\mathrm{X}_i)$ is the conditional variance of Y_i given X_i.

Proof. The CEF-decomposition property implies the variance of Y_i is the variance of the CEF plus the variance of the residual, $\varepsilon_i \equiv \mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]$ since ε_i and $E[\mathrm{Y}_i|\mathrm{X}_i]$ are uncorrelated. The variance of ε_i is

$$E[\varepsilon_i^2] = E[E[\varepsilon_i^2|\mathrm{X}_i]] = E[V[\mathrm{Y}_i|\mathrm{X}_i]]$$

where $E[\varepsilon_i^2|\mathrm{X}_i] = V[\mathrm{Y}_i|\mathrm{X}_i]$ because $\varepsilon_i \equiv \mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]$. ■

The two CEF properties and the ANOVA theorem may have a familiar ring. You might be used to seeing an ANOVA table in your regression output, for example. ANOVA is also important in research on inequality where labor economists decompose changes in the income distribution into parts that can be accounted for by changes in worker characteristics and changes in what's left over after accounting for these factors (See, e.g., Autor, Katz, and Kearney, 2005). What may be unfamiliar is the fact that the CEF properties and ANOVA variance decomposition work in the population as well as in samples, and do not turn on the assumption of a linear CEF. In fact, the validity of linear regression as an empirical tool does not turn on linearity either.

3.1.2 Linear Regression and the CEF

So what's the regression you want to run?

In our world, this question or one like it is heard almost every day. Regression estimates provide a valuable baseline for almost all empirical research because regression is tightly linked to the CEF, and the CEF

provides a natural summary of empirical relationships. The link between regression functions – i.e., the best-fitting line generated by minimizing expected squared errors – and the CEF can be explained in at least 3 ways. To lay out these explanations precisely, it helps to be precise about the regression function we have in mind. This chapter is concerned with the vector of *population* regression coefficients, defined as the solution to a population least squares problem. At this point, we are not worried about causality. Rather, we let the $\mathrm{K}\times 1$ regression coefficient vector β be defined by solving

$$\beta = \arg\min_{b} E\left[\left(\mathrm{Y}_i - \mathrm{X}_i'b\right)^2\right]. \tag{3.1.2}$$

Using the first-order condition,

$$E\left[\mathrm{X}_i\left(\mathrm{Y}_i - \mathrm{X}_i'b\right)\right] = 0.$$

the solution for b can be written $\beta = E\left[\mathrm{X}_i\mathrm{X}_i'\right]^{-1} E\left[\mathrm{X}_i\mathrm{Y}_i\right]$. Note that by construction, $E\left[\mathrm{X}_i\left(\mathrm{Y}_i - \mathrm{X}_i'\beta\right)\right] = 0$. In other words, the population residual, which we *define* as $\mathrm{Y}_i - \mathrm{X}_i'\beta = e_i$, is uncorrelated with the regressors, X_i. It bears emphasizing that this error term does not have a life of its own. It owes its existence and meaning to β.

In the simple bivariate case where the regression vector includes only the single regressor, x_i, and a constant, the slope coefficient is $\beta_1 = \frac{Cov(\mathrm{Y}_i, x_i)}{V(x_i)}$, and the intercept is $\alpha = E\left[\mathrm{Y}_i\right] - \beta_1 E\left[\mathrm{X}_i\right]$. In the multivariate case, i.e., with more than one non-constant regressor, the slope coefficient for the k-th regressor is given below:

REGRESSION ANATOMY

$$\beta_k = \frac{Cov\left(\mathrm{Y}_i, \tilde{x}_{ki}\right)}{V\left(\tilde{x}_{ki}\right)}, \tag{3.1.3}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all the other covariates.

In other words, $E\left[\mathrm{X}_i\mathrm{X}_i'\right]^{-1} E\left[\mathrm{X}_i\mathrm{Y}_i\right]$ is the $\mathrm{K}\times 1$ vector with k-th element $\frac{Cov(\mathrm{Y}_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}$. This important formula is said to describe the "anatomy of a multivariate regression coefficient" because it reveals much more than the matrix formula $\beta = E\left[\mathrm{X}_i\mathrm{X}_i'\right]^{-1} E\left[\mathrm{X}_i\mathrm{Y}_i\right]$. It shows us that each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor, after "partialling out" all the other variables in the model.

To verify the regression-anatomy formula, substitute

$$\mathrm{Y}_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + ... + \beta_{\mathrm{K}} x_{\mathrm{K}i} + e_i$$

in the numerator of (3.1.3). Since $\tilde{x}_{ki}$ is a linear combination of the regressors, it is uncorrelated with e_i. Also, since $\tilde{x}_{ki}$ is a residual from a regression on all the other covariates in the model, it must be uncorrelated these covariates. Finally, for the same reason, the covariance of $\tilde{x}_{ki}$ with x_{ki} is just the variance of $\tilde{x}_{ki}$. We

therefore have that $Cov\left(\mathrm{Y}_i, \tilde{x}_{ki}\right) = \beta_k V\left(\tilde{x}_{ki}\right)$.[2]

The regression-anatomy formula is probably familiar to you from a regression or statistics course, perhaps with one twist: the regression coefficients defined in this section are not estimators, but rather they are non-stochastic features of the joint distribution of dependent and independent variables. The joint distribution is what you would observe if you had a complete enumeration of the population of interest (or knew the stochastic process generating the data). You probably don't have such information. Still, it's kosher—even desirable—to think about what a set of population parameters might mean, without initially worrying about how to estimate them.

Below we discuss three reasons why the vector of population regression coefficients might be of interest. These reasons can be summarized by saying that you are interested in regression parameters if you are interested in the CEF.

Theorem 3.1.4 *The Linear CEF Theorem (Regression-justification I)*

Suppose the CEF is linear. Then the population regression function is it.

Proof. Suppose $E\left[\mathrm{Y}_i|\mathrm{X}_i\right] = \mathrm{X}_i'\beta^*$ for a $\mathrm{K}\times 1$ vector of coefficients, β^*. Recall that $E\left[\mathrm{X}_i\left(\mathrm{Y}_i - E\left[\mathrm{Y}_i|\mathrm{X}_i\right]\right)\right] = 0$ by the CEF-decomposition property. Substitute using $E\left[\mathrm{Y}_i|\mathrm{X}_i\right] = \mathrm{X}_i'\beta^*$ to find that $\beta^* = E\left[\mathrm{X}_i\mathrm{X}_i'\right]^{-1} E\left[\mathrm{X}_i\mathrm{Y}_i\right] = \beta$. ■

The linear CEF theorem raises the question of under what circumstances a CEF is linear. The classic scenario is joint Normality, i.e., the vector $\left(\mathrm{Y}_i, x_i'\right)'$ has a multivariate Normal distribution. This is the scenario considered by Galton (1886), father of regression, who was interested in the intergenerational link between Normally distributed traits such as height and intelligence. The Normal case is clearly of limited empirical relevance since regressors and dependent variables are often discrete, while Normal distributions are continuous. Another linearity scenario arises when regression models are saturated. As reviewed in Section 3.1.4, the saturated regression model has a separate parameter for every possible combination of values that the set of regressors can take on. For example a saturated regression model with two dummy covariates includes both covariates (with coefficients known as the main effects) and their product (known as an interaction term). Such models are inherently linear, a point we also discuss in Section 3.1.4.

[2] The regression-anatomy formula is usually attributed to Frisch and Waugh (1933). You can also do regression anatomy this way:

$$\beta_k = \frac{Cov\left(\tilde{\mathrm{Y}}_{ki}, \tilde{x}_{ki}\right)}{V\left(\tilde{x}_{ki}\right)},$$

where $\tilde{\mathrm{Y}}_{ki}$ is the residual from a regression of Y_i on every covariate except x_{ki}. This works because the fitted values removed from $\tilde{\mathrm{Y}}_{ki}$ are uncorrelated with $\tilde{x}_{ki}$. Often it's useful to plot $\tilde{\mathrm{Y}}_{ki}$ against $\tilde{x}_{ki}$; the slope of the least-squares fit in this scatterplot is your estimate of the multivariate β_k, even though the plot is two-dimensional. Note, however, that it's not enough to partial the other covariates out of Y_i only. That is,

$$\frac{Cov\left(\tilde{\mathrm{Y}}_{ki}, x_{ki}\right)}{V\left(x_{ki}\right)} = \left[\frac{Cov\left(\tilde{\mathrm{Y}}_{ki}, \tilde{x}_{ki}\right)}{V\left(\tilde{x}_{ki}\right)}\right]\left[\frac{V\left(\tilde{x}_{ki}\right)}{V\left(x_{ki}\right)}\right] \neq \beta_k,$$

unless x_{ki} is uncorrelated with the other covariates.

The following two reasons for focusing on regression are relevant when the linear CEF theorem does not apply.

Theorem 3.1.5 *The Best Linear Predictor Theorem (Regression-justification II)*

The function $\mathrm{X}_i'\beta$ *is the best* linear *predictor of* Y_i *given* X_i *in a MMSE sense.*

Proof. $\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$ solves the population least squares problem, (3.1.2). ■

In other words, just as the CEF, $E[\mathrm{Y}_i|\mathrm{X}_i]$, is the best (i.e., MMSE) predictor of Y_i given X_i in the class of *all* functions of X_i, the population regression function is the best we can do in the class of *linear* functions.

Theorem 3.1.6 *The Regression-CEF Theorem (Regression-justification III)*

The function $\mathrm{X}_i'\beta$ *provides the MMSE linear approximation to* $E[\mathrm{Y}_i|\mathrm{X}_i]$*, that is,*

$$\beta = \arg\min_b E\{(E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'b)^2\}. \tag{3.1.4}$$

Proof. Write

$$\begin{aligned}\left(\mathrm{Y}_i - \mathrm{X}_i'b\right)^2 &= \{(\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]) + (E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'b)\}^2 \\ &= (\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i])^2 + (E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'b)^2 \\ &\quad +2(\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i])(E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'b).\end{aligned}$$

The first term doesn't involve b and the last term has expectation zero by the CEF-decomposition property (ii). The CEF-approximation problem, (3.1.4), therefore has the same solution as the population least squares problem, (3.1.2). ■

These two theorems show us two more ways to view regression. Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable. On the other hand, if we prefer to think about approximating $E[\mathrm{Y}_i|\mathrm{X}_i]$, as opposed to predicting Y_i, the Regression-CEF theorem tells us that even if the CEF is nonlinear, regression provides the best linear approximation to it.

The regression-CEF theorem is our favorite way to motivate regression. The statement that regression approximates the CEF lines up with our view of empirical work as an effort to describe the essential features of statistical relationships, without necessarily trying to pin them down exactly. The linear CEF theorem is for special cases only. The best linear predictor theorem is satisfyingly general, but it encourages an overly clinical view of empirical research. We're not really interested in predicting *individual* Y_i; it's the *distribution* of Y_i that we care about.

Figure 3.1.2 illustrates the CEF approximation property for the same schooling CEF plotted in Figure 3.1.1. The regression line fits the somewhat bumpy and nonlinear CEF as if we were estimating a model

for $E[\mathrm{Y}_i|\mathrm{X}_i]$ instead of a model for Y_i. In fact, that is exactly what's going on. An implication of the regression-CEF theorem is that regression coefficients can be obtained by using $E[\mathrm{Y}_i|\mathrm{X}_i]$ as a dependent variable instead of Y_i itself. To see this, suppose that X_i is a discrete random variable with probability mass function, $g_x(u)$ when $\mathrm{X}_i = u$. Then

$$E\{(E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'b)^2\} = \sum_u (E[\mathrm{Y}_i|\mathrm{X}_i = u] - u'b)^2 g_x(u).$$

This means that β can be constructed from the weighted least squares regression of $E[\mathrm{Y}_i|\mathrm{X}_i = u]$ on u, where u runs over the values taken on by X_i. The weights are given by the distribution of X_i, i.e., $g_x(u)$ when $\mathrm{X}_i = u$. Another way to see this is to iterate expectations in the formula for β:

$$\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i] = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i E(\mathrm{Y}_i|\mathrm{X}_i)]. \tag{3.1.5}$$

The CEF or grouped-data version of the regression formula is of practical use when working on a project that precludes the analysis of micro data. For example, Angrist (1998), studies the effect of voluntary military service on earnings later in life. One of the estimation strategies used in this project regresses civilian earnings on a dummy for veteran status, along with personal characteristics and the variables used by the military to screen soldiers. The earnings data come from the US Social Security system, but Social Security earnings records cannot be released to the public. Instead of individual earnings, Angrist worked with average earnings conditional on race, sex, test scores, education, and veteran status.

An illustration of the grouped-data approach to regression appears below. We estimated the schooling coefficient in a wage equation using 21 conditional means, the sample CEF of earnings given schooling. As the Stata output reported here shows, a grouped-data regression, weighted by the number of individuals at each schooling level in the sample, produces coefficients *identical* to what would be obtained using the underlying microdata sample with hundreds of thousands of observations. Note, however, that the standard errors from the grouped regression do not correctly reflect the asymptotic sampling variance of the slope estimate in repeated *micro-data* samples; for that you need an estimate of the variance of $\mathrm{Y}_i - \mathrm{X}_i'\beta$. This variance depends on the microdata, in particular, the second-moments of $W_i \equiv \left[\begin{array}{cc} \mathrm{Y}_i; & \mathrm{X}_i' \end{array}\right]'$, a point we elaborate on in the next section.

3.1.3 Asymptotic OLS Inference

In practice, we don't usually know what the CEF or the population regression vector is. We therefore draw statistical inferences about these quantities using samples. Statistical inference is what much of traditional econometrics is about. Although this material is covered in any Econometrics text, we don't want to skip the inference step completely. A review of basic asymptotic theory allows us to highlight the important fact that the process of statistical inference is entirely distinct from the question of how a particular set of regression

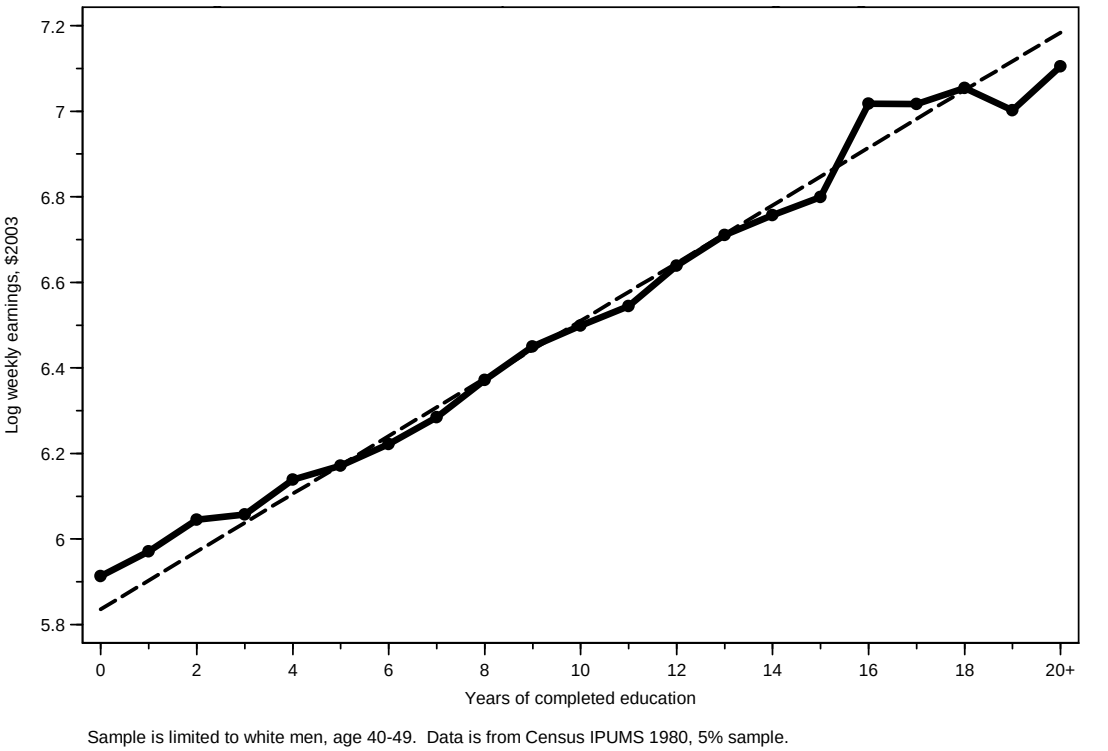

Figure 3.1.2: Regression threads the CEF of average weekly wages given schooling

estimates should be interpreted. Whatever a regression coefficient may mean, it has a sampling distribution that is easy to describe and use for statistical inference.[3]

We are interested in the distribution of the sample analog of

$$\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$$

in repeated samples. Suppose the vector $W_i \equiv \left[\begin{array}{cc} \mathrm{Y}_i; & \mathrm{X}_i' \end{array} \right]'$ is independently and identically distributed in a sample of size N. A natural estimator of the first population moment, $E[W_i]$, is the sum, $\frac{1}{N}\sum_{i=1}^{N} W_i$. By the law of large numbers, this sample moment gets arbitrarily close to the corresponding population moment as the sample size grows. We might similarly consider higher-order moments of the elements of W_i, e.g., the matrix of second moments, $E[W_iW_i']$, with sample analog $\frac{1}{N}\sum_{i=1}^{N} W_iW_i'$. Following this principle, the method of moments estimator of β replaces each expectation by a sum. This logic leads to the Ordinary Least Squares (OLS) estimator

$$\hat{\beta} = \left[\sum_i \mathrm{X}_i\mathrm{X}_i'\right]^{-1} \sum_i \mathrm{X}_i\mathrm{Y}_i.$$

Although we derived $\hat{\beta}$ as a method of moments estimator, it is called the OLS estimator of β because it solves the sample analog of the least-squares problem described at the beginning of Section 3.1.2.[4]

[3] The discussion of asymptotic OLS inference in this section is largely a condensation of material in Chamberlain (1984). Important pitfalls and problems with this asymptotic theory are covered in the last chapter.

[4] Econometricians like to use matrices because the notation is so compact. Sometimes (not very often) we do too. Suppose

A - Individual-level data

```
. regress earnings school, robust

      Source |       SS       df       MS        Number of obs =  409435
-------------+------------------------------     F(  1,409433) =49118.25
       Model | 22631.4793      1  22631.4793     Prob > F      =  0.0000
    Residual |  188648.31 409433  .460755019     R-squared     =  0.1071
-------------+------------------------------     Adj R-squared =  0.1071
       Total | 211279.789 409434   .51602893     Root MSE      =  .67879
-------------+----------------------------------------------------------
             |               Robust                 Old Fashioned
    earnings |      Coef.   Std. Err.      t           Std. Err.       t
-------------+----------------------------------------------------------
      school |   .0674387   .0003447   195.63          .0003043   221.63
      const. |   5.835761   .0045507  1282.39          .0040043  1457.38
------------------------------------------------------------------------
```

B - Means by years of schooling

```
. regress average_earnings school [aweight=count], robust
(sum of wgt is   4.0944e+05)

      Source |       SS       df       MS        Number of obs =      21
-------------+------------------------------     F(  1,    19) =  540.31
       Model |  1.16077332     1  1.16077332     Prob > F      =  0.0000
    Residual |  .040818796    19  .002148358     R-squared     =  0.9660
-------------+------------------------------     Adj R-squared =  0.9642
       Total |  1.20159212    20  .060079606     Root MSE      =  .04635
-------------+----------------------------------------------------------
     average |               Robust                 Old Fashioned
   _earnings |      Coef.   Std. Err.      t           Std. Err.       t
-------------+----------------------------------------------------------
      school |   .0674387   .0040352    16.71          .0029013     23.24
      const. |   5.835761   .0399452   146.09          .0381792    152.85
------------------------------------------------------------------------
```

Figure 3.1.3: Micro-data and grouped-data estimates of returns to schooling. Source: 1980 Census - IPUMS, 5 percent sample. Sample is limited to white men, age 40-49. Derived from Stata regression output. Old-fashioned standard errors are the default reported. Robust standard errors are heteroscedasticity-consistent. Panel A uses individual-level data. Panel B uses earnings averaged by years of schooling.

The asymptotic sampling distribution of $\hat{\beta}$ depends solely on the definition of the estimand (i.e., the nature of the thing we're trying to estimate, β) and the assumption that the data constitute a random sample. Before deriving this distribution, it helps to record the general asymptotic distribution theory that covers our needs. This basic theory can be stated mostly in words. For the purposes of these statements, we assume the reader is familiar with the core terms and concepts of statistical theory (e.g., moments, mathematical expectation, probability limits, and asymptotic distributions). For definitions of these terms and a formal mathematical statement of the theoretical propositions given below, see, e.g., Knight (2000).

THE LAW OF LARGE NUMBERS Sample moments converge in probability to the corresponding population moments. In other words, the probability that the sample mean is close to the population mean can be made as high as you like by taking a large enough sample.

THE CENTRAL LIMIT THEOREM Sample moments are asymptotically Normally distributed (after subtracting the corresponding population moment and multiplying by the square root of the sample size). The covariance matrix is given by the variance of the underlying random variable. In other words, in large enough samples, appropriately normalized sample moments are approximately Normally distributed.

SLUTSKY'S THEOREM

(a) Consider the sum of two random variables, one of which converges in distribution and the other converges in probability to a constant: the asymptotic distribution of this sum is unaffected by replacing the one that converges to a constant by this constant. Formally, let a_N be a statistic with a limiting distribution and let b_N be a statistic with probability limit b. Then $a_N + b_N$ and $a_N + b$ have the same limiting distribution.

(b) Consider the product of two random variables, one of which converges in distribution and the other converges in probability to a constant: the asymptotic distribution of this product is unaffected by replacing the one that converges to a constant by this constant. This allows us to replaces some sample moments by population moments (i.e., by their probability limits) when deriving distributions. Formally, let a_N be a statistic with a limiting distribution and let b_N be a statistic with probability limit b. Then $a_N b_N$ and $a_N b$ have the same asymptotic distribution.

THE CONTINUOUS MAPPING THEOREM Probability limits pass through continuous functions. For example, the probability limit of any continuous function of a sample moment is the function evaluated at the corresponding population moment. Formally, the probability limit of $h(b_N)$ is $h(b)$ where $plim\ b_N = b$ and $h(\cdot)$ is continuous at b.

X is the matrix whose rows are given by X_i' and y is the vector with elements y_i, for $i = 1, ..., N$. The sample moment $\frac{1}{N}\sum X_i X_i'$ is $X'X/N$ and the sample moment $\frac{1}{N}\sum X_i y_i$ is $X'y/N$. Then we can write $\hat{\beta} = (X'X)^{-1}X'y$, a familiar matrix formula.

THE DELTA METHOD Consider a vector-valued random variable that is asymptotically Normally distributed. Most scalar functions of this random variable are also asymptotically Normally distributed, with covariance matrix given by a quadratic form with the covariance matrix of the random variable on the inside and the gradient of the function evaluated at the probability limit of the random variable on the outside. Formally, the asymptotic distribution of $h(b_N)$ is Normal with covariance matrix $\nabla h(b)'\Omega\nabla h(b)$ where $plim\ b_N = b$, $h(\cdot)$ is continuously differentiable at b with gradient $\nabla h(b)$, and b_N has asymptotic covariance matrix Ω.[5]

We can use these results to derive the asymptotic distribution of $\hat{\beta}$ in two ways. A conceptually straightforward but somewhat inelegant approach is to use the delta method: $\hat{\beta}$ is a function of sample moments, and is therefore asymptotically Normally distributed. It remains only to find the covariance matrix of the asymptotic distribution from the gradient of this function. (Note that consistency of $\hat{\beta}$ comes immediately from the continuous mapping theorem). An easier and more instructive derivation uses the Slutsky and central limit theorems. Note first that we can write

$$\mathrm{Y}_i = \mathrm{X}_i'\beta + [\mathrm{Y}_i - \mathrm{X}_i'\beta] \equiv \mathrm{X}_i'\beta + e_i, \tag{3.1.6}$$

where the residual e_i is *defined* as the difference between the dependent variable and the population regression function, as before. This is as good a place as any to point out that these residuals are uncorrelated with the regressors *by definition of* β. In other words, $E[\mathrm{X}_i e_i] = 0$ is a consequence of $\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$ and $e_i = \mathrm{Y}_i - \mathrm{X}_i'\beta$, and not an assumption about an underlying economic relation. We return to this important point in the discussion of causal regression models in Section 3.2.[6]

Substituting the identity 3.1.6 for Y_i in the formula for $\hat{\beta}$, we have

$$\hat{\beta} = \beta + \left[\sum \mathrm{X}_i\mathrm{X}_i'\right]^{-1}\sum \mathrm{X}_i e_i.$$

The asymptotic distribution of $\hat{\beta}$ is the asymptotic distribution of $\sqrt{N}(\hat{\beta}-\beta) = N\left[\sum \mathrm{X}_i\mathrm{X}_i'\right]^{-1}\frac{1}{\sqrt{N}}\sum \mathrm{X}_i e_i$. By the Slutsky theorem, this has the same asymptotic distribution as $E[\mathrm{X}_i\mathrm{X}_i']^{-1}\frac{1}{\sqrt{N}}\sum \mathrm{X}_i e_i$. Since $E[\mathrm{X}_i e_i] = 0$, $\frac{1}{\sqrt{N}}\sum \mathrm{X}_i e_i$ is a root-N-normalized and centered sample moment. By the central limit theorem, this is asymptotically Normally distributed with mean zero and covariance matrix $E[\mathrm{X}_i\mathrm{X}_i'e_i^2]$, since this fourth moment is the covariance matrix of $\mathrm{X}_i e_i$. Therefore, $\hat{\beta}$ has an asymptotic Normal distribution, with probability limit β, and covariance matrix

$$E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{X}_i'e_i^2]E[\mathrm{X}_i\mathrm{X}_i']^{-1}. \tag{3.1.7}$$

The standard errors used to construct t-statistics are the square roots of the diagonal elements of this

[5] For a derivation of the the delta method formula using the Slutsky and continuous mapping theorems, see, e.g., Knight, 2000, pp. 120-121.

[6] Residuals defined in this way are not necessarily *mean-independent* of X_i; for mean-independence, we need a linear CEF.

matrix. In practice these standard errors are estimated by substituting sums for expectations, and using the estimated residuals, $\hat{e}_i = \mathrm{Y}_i - \mathrm{X}_i'\hat{\beta}$ to form the empirical fourth moment, $\sum[\mathrm{X}_i\mathrm{X}_i\hat{e}_i^2]/N$.

Asymptotic standard errors computed in this way are known as heteroskedasticity-consistent standard errors, White (1980a) standard errors, or Eicker-White standard errors in recognition of Eicker's (1967) derivation. They are also known as "robust" standard errors (e.g., in Stata). These standard errors are said to be robust because, in large enough samples, they provide accurate hypothesis tests and confidence intervals given minimal assumptions about the data and model. In particular, our derivation of the limiting distribution makes no assumptions other than those needed to ensure that basic statistical results like the central limit theorem go through. These are not, however, the standard errors that you get by default from packaged software. Default standard errors are derived under a homoskedasticity assumption, specifically, that $E[e_i^2|\mathrm{X}_i] = \sigma^2$, a constant. Given this assumption, we have

$$E[\mathrm{X}_i\mathrm{X}_i'e_i^2] = E(\mathrm{X}_i\mathrm{X}_i'E[e_i^2|\mathrm{X}_i]) = \sigma^2 E[\mathrm{X}_i\mathrm{X}_i'],$$

by iterating expectations. The asymptotic covariance matrix of $\hat{\beta}$ then simplifies to

$$\begin{aligned} E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{X}_i'e_i^2]E[\mathrm{X}_i\mathrm{X}_i']^{-1} &= E[\mathrm{X}_i\mathrm{X}_i']^{-1}\sigma^2 E[\mathrm{X}_i\mathrm{X}_i']E[\mathrm{X}_i\mathrm{X}_i]^{-1} \\ &= E[\mathrm{X}_i\mathrm{X}_i']^{-1}\sigma^2. \end{aligned} \tag{3.1.8}$$

The diagonal elements of (3.1.8) are what SAS or Stata report unless you request otherwise.

Our view of regression as an approximation to the CEF makes heteroskedasticity seem natural. If the CEF is nonlinear and you use a linear model to approximate it, then the quality of fit between the regression line and the CEF will vary with X_i. Hence, the residuals will be larger, on average, at values of X_i where the fit is poorer. Even if you are prepared to assumed that the conditional variance of Y_i given X_i is constant, the fact that the CEF is nonlinear means that $E[(\mathrm{Y}_i - \mathrm{X}_i'\beta)^2|\mathrm{X}_i]$ will vary with X_i. To see this, note that, as a rule,

$$\begin{aligned} E[(\mathrm{Y}_i - \mathrm{X}_i'\beta)^2|\mathrm{X}_i] &= \\ &\quad E\{[(\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]) + (E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'\beta)]^2|\mathrm{X}_i\} \\ &= V[\mathrm{Y}_i|\mathrm{X}_i] + (E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'\beta)^2. \end{aligned} \tag{3.1.9}$$

Therefore, even if $V[\mathrm{Y}_i|\mathrm{X}_i]$ is constant, the residual variance increases with the square of the gap between the regression line and the CEF, a fact noted in White (1980b).[7]

In the same spirit, it's also worth noting that while a linear CEF makes homoskedasticity possible, this is

[7] The cross-product term resulting from an expansion of the quadratic in the middle of 3.1.9 is zero because $\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]$ is mean-independent of X_i.

not a sufficient condition for homoskedasticity. Our favorite example in this context is the linear probability model (LPM). A linear probability model is any regression where the dependent variable is zero-one, i.e., a dummy variable such as an indicator for labor force participation. Suppose the regression model is saturated, so the CEF is linear. Because the CEF is linear, the residual variance is also the conditional variance, $V[\mathrm{Y}_i|\mathrm{X}_i]$. But the dependent variable is a Bernoulli trial and the variance of a Bernoulli trial is $P[\mathrm{Y}_i|\mathrm{X}_i](1 - P[\mathrm{Y}_i|\mathrm{X}_i])$. We conclude that LPM residuals are necessarily heteroskedastic unless the only regressor is a constant.

These points of principle notwithstanding, as an empirical matter, heteroskedasticity may matter little. In the micro-data schooling regression depicted in Figure 3.1.3, the robust standard error is .0003447, while the old-fashioned standard error is .0003043, only slightly smaller. The standard errors from the grouped-data regression, which are necessarily heteroskedastic if group sizes differ, change somewhat more; compare the .004 robust standard to the .0029 conventional standard error. Based on our experience, these differences are typical. If heteroskedasticity matters too much, say, more than a 30% increase or any marked decrease in standard errors, you should worry about possible programming errors or other problems (for example, robust standard errors below conventional may be a sign of finite-sample bias in the robust calculation; see Chapter 8, below.)

3.1.4 Saturated Models, Main Effects, and Other Regression Talk

We often discuss regression models using terms like *saturated* and *main effects*. These terms originate in an experimentalist tradition that uses regression to model discrete treatment-type variables. This language is now used more widely in many fields, however, including applied econometrics. For readers unfamiliar with these terms, this section provides a brief review.

Saturated regression models are regression models with discrete explanatory variables, where the model includes a separate parameter for all possible values taken on by the explanatory variables. For example, when working with a single explanatory variable indicating whether a worker is a college graduate, the model is saturated by including a single dummy for college graduates and a constant. We can also saturate when the regressor takes on many values. Suppose, for example, that $\mathrm{S}_i = 0, 1, 2, ..., \tau$. A saturated regression model for S_i is

$$\mathrm{Y}_i = \beta_0 + \beta_1 d_{1i} + \beta_2 d_{2i} + ... + \beta_\tau d_{\tau i} + \varepsilon_i,$$

where $d_{ji} = 1[\mathrm{S}_i = j]$ is a dummy variable indicating schooling level-j, and β_j is said to be the jth-level schooling *effect*. Note that

$$\beta_j = E[\mathrm{Y}_i|\mathrm{S}_i = j] - E[\mathrm{Y}_i|\mathrm{S}_i = 0],$$

while $\beta_0 = E[\mathrm{Y}_i|\mathrm{S}_i = 0]$. In practice, you can pick any value of S_i for the reference group; a regression model is saturated as long as it has one parameter for every possible j in $E[\mathrm{Y}_i|\mathrm{S}_i = j]$. Saturated models fit the

CEF perfectly because the CEF is linear in the dummy regressors used to saturate. This is an important special case of the regression-CEF theorem.

If there are two explanatory variables, say one dummy indicating college graduates and one dummy indicating sex, the model is saturated by including these two dummies, their product, and a constant. The coefficients on the dummies are known as main effects, while the product is called an *interaction term*. This is not the only saturated parameterization; any set of indicators (dummies) that can be used to identify each value taken on by the covariates produces a saturated model. For example, an alternative saturated model includes dummies for male college graduates, male dropouts, female college graduates, and female dropouts, but no intercept.

Here's some notation to make this more concrete. Let x_{1i} indicate college graduates and x_{2i} indicate women. The CEF given x_{1i} and x_{2i} takes on four values:

$$\begin{aligned} &E\left[\mathrm{Y}_i | x_{1i}=0, x_{2i}=0\right], \\ &E\left[\mathrm{Y}_i | x_{1i}=1, x_{2i}=0\right], \\ &E\left[\mathrm{Y}_i | x_{1i}=0, x_{2i}=1\right], \\ &E\left[\mathrm{Y}_i | x_{1i}=1, x_{2i}=1\right]. \end{aligned}$$

We can label these using the following scheme:

$$\begin{aligned} E\left[\mathrm{Y}_i | x_{1i}=0, x_{2i}=0\right] &= \alpha \\ E\left[\mathrm{Y}_i | x_{1i}=1, x_{2i}=0\right] &= \alpha+\beta \\ E\left[\mathrm{Y}_i | x_{1i}=0, x_{2i}=1\right] &= \alpha+\gamma \\ E\left[\mathrm{Y}_i | x_{1i}=1, x_{2i}=1\right] &= \alpha+\beta+\gamma+\delta. \end{aligned}$$

Since there are four Greek letters and the CEF takes on four values, this parameterization does not restrict the CEF. It can be written in terms of Greek letters as

$$E[\mathrm{Y}_i | x_{1i}, x_{2i}] = \alpha + \beta x_{1i} + \gamma x_{2i} + \delta(x_{1i} x_{2i}),$$

a parameterization with two main effects and one interaction term.[8] The saturated regression equation becomes

$$\mathrm{Y}_i = \alpha + \beta x_{1i} + \gamma x_{2i} + \delta(x_{1i} x_{2i}) + \varepsilon_i.$$

Finally, we can combine the multi-valued schooling variable with sex to produce a saturated model that

[8] With a third dummy variable in the model, say x_{3i}, a saturated model includes 3 main effects, 3 second-order interaction terms $\{x_{1i}x_{2i},\ x_{2i}x_{3i}, x_{1i}x_{2i}\}$ and one third-order term, $x_{1i}x_{2i}x_{3i}$.

has τ main effects for schooling, one main effect for sex, and τ sex-schooling interactions:

$$\text{Y}_i = \beta_0 + \sum_{j=1}^{\tau} \beta_j d_{ji} + \gamma x_{2i} + \sum_{j=1}^{\tau} \delta_j (d_{ji} x_{2i}) + \varepsilon_i. \tag{3.1.10}$$

The interaction terms, δ_j, tell us how each of the schooling effects differ by sex. The CEF in this case takes on $2(\tau + 1)$ values while the regression has this many parameters.

Note that there is a natural hierarchy of modeling strategies with saturated models at the top. It's natural to start with a saturated model because this fits the CEF. On the other hand, saturated models generate a lot of interaction terms, many of which may be uninteresting or imprecise. You might therefore sensibly choose to omit some or all of these. Equation (3.1.10) without interaction terms approximates the CEF with a purely additive model for schooling and sex. This is a good approximation if the returns to college are similar for men and women. And, in any case, schooling coefficients in the additive specification give a (weighted) average return across both sexes, as discussed in Section 3.3.1, below. On the other hand, it would be strange to estimate a model which included interaction terms but omitted the corresponding main effects. In the case of schooling, this would be something like

$$\text{Y}_i = \beta_0 + \gamma x_{2i} + \sum_{j=1}^{\tau} \delta_j (d_{ji} x_{2i}) + \varepsilon_i. \tag{3.1.11}$$

This model allows schooling to shift wages only for women, something very far from the truth. Consequently, the results of estimating (3.1.11) are likely to be hard to interpret.

Finally, it's important to recognize that a saturated model fits the CEF perfectly regardless of the distribution of Y_i. For example, this is true for linear probability models and other limited dependent variable models (e.g., non-negative Y_i), a point we return to at the end of this chapter.

3.2 Regression and Causality

Section 3.1.2 shows how regression gives the best (MMSE) linear approximation to the CEF. This understanding, however, does not help us with the deeper question of when regression has a causal interpretation. When can we think of a regression coefficient as approximating the causal effect that might be revealed in an experiment?

3.2.1 The Conditional Independence Assumption

A regression is causal when the CEF it approximates is causal. This doesn't answer the question, of course. It just passes the buck up one level, since, as we've seen, a regression inherits it's legitimacy from a CEF. Causality means different things to different people, but researchers working in many disciplines have found it useful to think of causal relationships in terms of the potential outcomes notation used in Chapter 2 to

describe what would happen to a given individual in a hypothetical comparison of alternative hospitalization scenarios. Differences in these potential outcomes were said to be the causal effect of hospitalization. The CEF is causal when it describes differences in average potential outcomes for a fixed reference population.

It's easiest to expand on the somewhat murky notion of a causal CEF in the context of a particular question, so let's stick with the schooling example. The causal connection between schooling and earnings can be defined as the functional relationship that describes what a given individual would earn if he or she obtained different levels of education. In particular, we might think of schooling decisions as being made in a series of episodes where the decision-maker might realistically go one way or another, even if certain choices are more likely than others. For example, in the middle of junior year, restless and unhappy, Angrist glumly considered his options: dropping out of high school and hopefully getting a job, staying in school but taking easy classes that lead to a quick and dirty high school diploma, or plowing on in an academic track that leads to college. Although the consequences of such choices are usually unknown in advance, the idea of alternative paths leading to alternative outcomes for a given individual seems uncontroversial. Philosophers have argued over whether this personal notion of potential outcomes is precise enough to be scientifically useful, but individual decision-makers seem to have no trouble thinking about their lives and choices in this manner (as in Robert Frost's celebrated *The Road Not Taken*: the traveller-narrator sees himself looking back on a moment of choice. He believes that the decision to follow the road less traveled "has made all the difference," though he also recognizes that counterfactual outcomes are unknowable).

In empirical work, the causal relationship between schooling and earnings tells us what people would earn—on average—if we could either change their schooling in a perfectly-controlled environment, or change their schooling randomly so that those with different levels of schooling would be otherwise comparable. As we discussed in Chapter 2, experiments ensure that the causal variable of interest is independent of potential outcomes so that the groups being compared are truly comparable. Here, we would like to generalize this notion to causal variables that take on more than two values, and to more complicated situations where we must hold a variety of "control variables" fixed for causal inferences to be valid. This leads to the *conditional independence assumption* (CIA), a core assumption that provides the (sometimes implicit) justification for the causal interpretation of regression. This assumption is sometimes called selection-on-observables because the covariates to be held fixed are assumed to be known and observed (e.g., in Goldberger, 1972; Barnow, Cain, and Goldberger, 1981). The big question, therefore, is what these control variables are, or should be. We'll say more about that shortly. For now, we just do the econometric thing and call the covariates "X_i". As far as the schooling problem goes, it seems natural to imagine that X_i is a vector that includes measures of ability and family background.

For starters, think of schooling as a binary decision, like whether Angrist goes to college. Denote this by a dummy variable, c_i. The causal relationship between college attendance and a future outcome like earnings can be described using the same potential-outcomes notation we used to describe experiments in

Chapter 2. To address this question, we imagine two potential earnings variables:

$$potential\ outcome = \begin{cases} \mathrm{Y}_{1i} & \text{if } \mathrm{C}_i = 1 \\ \mathrm{Y}_{0i} & \text{if } \mathrm{C}_i = 0 \end{cases}.$$

In this case, Y_{0i} is i's earnings without college, while Y_{1i} is i's earnings if he goes. We would like to know the difference between Y_{1i} and Y_{0i}, which is the causal effect of college attendance on individual i. This is what we would measure if we could go back in time and nudge i onto the road not taken. The observed outcome, Y_i, can be written in terms of potential outcomes as

$$\mathrm{Y}_i = \mathrm{Y}_{0i} + (\mathrm{Y}_{1i} - \mathrm{Y}_{0i})\mathrm{C}_i.$$

We get to see one of Y_{1i} or Y_{0i}, but never both. We therefore hope to measure the average of $\mathrm{Y}_{1i} - \mathrm{Y}_{0i}$, or the average for some group, such as those who went to college. This is $E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | \mathrm{C}_i = 1]$.

In general, comparisons of those who do and don't go to college are likely to be a poor measure of the causal effect of college attendance. Following the logic in Chapter 2, we have

$$\underbrace{E\left[\mathrm{Y}_i | \mathrm{C}_i = 1\right] - E[\mathrm{Y}_i | \mathrm{C}_i = 0]}_{\text{Observed difference in earnings}} = \underbrace{E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | \mathrm{C}_i = 1]}_{\text{average treatment effect on the treated}} + \underbrace{E\left[\mathrm{Y}_{0i} | \mathrm{C}_i = 1\right] - E\left[\mathrm{Y}_{0i} | \mathrm{C}_i = 0\right]}_{\text{selection bias}}. \tag{3.2.1}$$

It seems likely that those who go to college would have earned more anyway. If so, selection bias is positive, and the naive comparison, $E\left[\mathrm{Y}_i | \mathrm{C}_i = 1\right] - E[\mathrm{Y}_i | \mathrm{C}_i = 0]$, exaggerates the benefits of college attendance.

The CIA asserts that conditional on observed characteristics, X_i, selection bias disappears. In this example, the CIA says,

$$\{\mathrm{Y}_{0i}, \mathrm{Y}_{1i}\} \perp\!\!\!\perp \mathrm{C}_i | \mathrm{X}_i. \tag{3.2.2}$$

Given the CIA, conditional-on-X_i comparisons of average earnings across schooling levels have a causal interpretation. In other words,

$$E\left[\mathrm{Y}_i | \mathrm{X}_i, \mathrm{C}_i = 1\right] - E\left[\mathrm{Y}_i | \mathrm{X}_i, \mathrm{C}_i = 0\right] = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | \mathrm{X}_i].$$

Now, we'd like to expand the conditional independence assumption to causal relations that involve variables that can take on more than two values, like years of schooling, S_i. The causal relationship between schooling and earnings is likely to be different for each person. We therefore use the individual-specific notation,

$$\mathrm{Y}_{si} \equiv f_i(s)$$

to denote the potential earnings that person i would receive after obtaining s years of education. If s takes on only two values, 12 and 16, then we are back to the college/no college example:

$$\mathrm{Y}_{0i} = f_i(12); \mathrm{Y}_{1i} = f_i(16).$$

More generally, the function $f_i(s)$ tells us what i would earn for *any* value of schooling, s. In other words, $f_i(s)$ answers causal "what if" questions. In the context of theoretical models of the relationship between human capital and earnings, the form of $f_i(s)$ may be determined by aspects of individual behavior and/or market forces.

The CIA in this more general setup becomes

$$Y_{si} \amalg \mathrm{S}_i | \mathrm{X}_i \tag{CIA}$$

In many randomized experiments, the CIA crops up because S_i *is* randomly assigned conditional on X_i (In the Tennessee STAR experiment, for example, small classes were randomly assigned within schools). In an observational study, the CIA means that S_i can be said to be "as good as randomly assigned," conditional on X_i.

Conditional on X_i, the *average causal effect* of a one year increase in schooling is $E[f_i(s) - f_i(s-1)|\mathrm{X}_i]$, while the average causal effect of a 4-year increase in schooling is $E[f_i(s) - E\left[f_i(s-4)\right]|\mathrm{X}_i]$. The data reveal only $\mathrm{Y}_i = f_i(\mathrm{S}_i)$, however, that is $f_i(s)$ for $s = \mathrm{S}_i$. But given the CIA, conditional-on-X_i comparisons of average earnings across schooling levels have a causal interpretation. In other words,

$$\begin{aligned} & E\left[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{S}_i = s\right] - E\left[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{S}_i = s-1\right] \\ = \; & E\left[f_i(s) - f_i(s-1)|\mathrm{X}_i\right] \end{aligned}$$

for any value of s. For example, we can compare the earnings of those with 12 and 11 years of schooling to learn about the average causal effect of high school graduation:

$$E\left[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{S}_i = 12\right] - E\left[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{S}_i = 11\right] = E\left[f_i(12)|\mathrm{X}_i, \mathrm{S}_i = 12\right] - E\left[f_i(11)|\mathrm{X}_i, \mathrm{S}_i = 11\right].$$

This comparison has a causal interpretation because, given the CIA,

$$E\left[f_i(12)|\mathrm{X}_i, \mathrm{S}_i = 12\right] - E\left[f_i(11)|\mathrm{X}_i, \mathrm{S}_i = 11\right] = E\left[f_i(12) - f_i(11)|\mathrm{X}_i, \mathrm{S}_i = 12\right].$$

Here, the selection bias term is the average difference in the potential dropout-earnings of high school graduates and dropouts. Given the CIA, however, high school graduation is independent of potential earnings conditional on X_i, so the selection-bias vanishes. Note also that in this case, the causal effect of

graduating high school on high school graduates is the population average high school graduation effect:

$$E\left[f_i(12) - f_i(11)|\mathrm{X}_i, \mathrm{S}_i = 12\right] = E\left[f_i(12) - f_i(11)|\mathrm{X}_i\right].$$

This is important . . . but less important than the elimination of selection bias in (3.2.1).

So far, we have constructed separate causal effects for each value taken on by the conditioning variable, X_i. This leads to as many causal effects as there are values of X_i, an embarrassment of riches. Empiricists almost always find it useful to boil a set of estimates down to a single summary measure, like the population average causal effect. By the law of iterated expectations, the population average causal effect of high school graduation is

$$\begin{aligned} & E\left\{E\left[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{S}_i = 12\right] - E\left[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{S}_i = 11\right]\right\} && (3.2.3) \\ = \; & E\left\{E\left[f_i(12) - f_i(11)|\mathrm{X}_i\right]\right\} \\ = \; & E\left[f_i(12) - f_i(11)\right] && (3.2.4) \end{aligned}$$

In the same spirit, we might be interested in the average causal effect of high school graduation on high school graduates:

$$\begin{aligned} & E\{E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{S}_i = 12] - E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{S}_i = 11]|\mathrm{S}_i = 12\} && (3.2.5) \\ = \; & E\{E[f_i(12) - f_i(11)|\mathrm{X}_i]|\mathrm{S}_i = 12\} \\ = \; & E[f_i(12) - f_i(11)|\mathrm{S}_i = 12]. && (3.2.6) \end{aligned}$$

This parameter tells us how much high school graduates gained by virtue of having graduated. Likewise, for the effects of college graduation there is a distinction between $E[f_i(16) - f_i(12)|\mathrm{S}_i = 16]$, the average causal effect on college graduates and $E[f_i(16) - f_i(12)]$, the population average effect.

The population average effect, (3.2.3), can be computed by averaging all of the X-specific effects using the marginal distribution of X_i, while the average effect on high school or college graduates averages the X-specific effects using the distribution of X_i in these groups. In both cases, the empirical counterpart is a matching estimator: we make comparisons across schooling groups graduates for individuals with the same covariate values, compute the difference in their earnings, and then average these differences in some way.

In practice, there are many details to worry about when implementing a matching strategy. We fill in some of the technical details on the mechanics of matching in Section 3.3.1, below. Here we note that a global drawback of the matching approach is that it is not "automatic," rather it requires two steps, matching and averaging. Estimating the standard errors of the resulting estimates may not be straightforward, either.

A third consideration is that the two-way contrast at the heart of this subsection (high school or college completers versus dropouts) does not do full justice to the problem at hand. Since S_i takes on many values, there are separate average causal effects for each possible increment in S_i, which also must be summarized in some way.[9] These considerations lead us back to regression.

Regression provides an easy-to-use empirical strategy that automatically turns the CIA into causal effects. Two routes can be traced from the CIA to regression. One assumes that $f_i(s)$ is both linear in s and the same for everyone except for an additive error term**,** in which case linear regression is a natural tool to estimate the features of $f_i(s)$. A more general but somewhat longer route recognizes that $f_i(s)$ almost certainly differs for different people, and, moreover, need not be linear in s. Even so, allowing for random variation in $f_i(s)$ across people, and for non-linearity for a given person, regression can be thought of as strategy for the estimation of a weighted average of the individual-specific difference, $f_i(s) - f_i(s-1)$. In fact, regression can be seen as a particular sort of matching estimator, capturing an average causal effect much like 3.2.3 or 3.2.5.

At this point, we want to focus on the conditions required for regression to have a causal interpretation and not on the details of the regression-matching analog. We therefore start with the first route, a linear constant-effects causal model. Suppose that

$$f_i(s) = \alpha + \rho s + \eta_i. \tag{3.2.7}$$

In addition to being linear, this equation says that the functional relationship of interest is the same for everyone. Again, s is written without an i subscript to index individuals, because equation (3.2.7) tells us what person i would earn for any value of s and not just the realized value, S_i. In this case, however, the only individual-specific and random part of $f_i(s)$ is a mean-zero error component, η_i, which captures unobserved factors that determine potential earnings.

Substituting the observed value S_i for s in equation (3.2.7), we have

$$Y_i = \alpha + \rho S_i + \eta_i. \tag{3.2.8}$$

Equation (3.2.8) looks like a bivariate regression model, except that equation (3.2.7) explicitly associates the coefficients in (3.2.8) with a causal relationship. Importantly, because equation (3.2.7) is a causal model, S_i may be correlated with potential outcomes, $f_i(s)$, or, in this case, the residual term in (3.2.8), η_i.

[9] For example, we might construct the average effect over s using the distribution of S_i. In other words, estimate $E[f_i(s) - f_i(s-1)]$ for each s by matching, and then compute the average difference

$$\sum E[f_i(s) - f_i(s-1)]P(s).$$

where $P(s)$ is the probability mass function for S_i. This is a discrete approximation to the average derivative, $E[f_i'(S_i)]$.

Suppose now that the CIA holds given a vector of observed covariates, X_i. In addition to the functional form assumption for potential outcomes embodied in (3.2.8), we decompose the random part of potential earnings, η_i, into a linear function of observable characteristics, X_i, and an error term, v_i:

$$\eta_i = \mathrm{X}_i'\gamma + v_i,$$

where γ is a vector of population regression coefficients that is assumed to satisfy $E[\eta_i|\mathrm{X}_i] = \mathrm{X}_i'\gamma$. Because γ is defined by the regression of η_i on X_i,the residual v_i and X_i are uncorrelated *by construction*. Moreover, by virtue of the CIA, we have

$$E[f_i(s)|\mathrm{X}_i, \mathrm{S}_i] = E[f_i(s)|\mathrm{X}_i] = \alpha + \rho s + E[\eta_i|\mathrm{X}] = \alpha + \rho s + \mathrm{X}_i'\gamma$$

Because mean-independence implies orthogonality, the residual in the linear causal model

$$\mathrm{Y}_i = \alpha + \rho \mathrm{S}_i + \mathrm{X}_i'\gamma + v_i \tag{3.2.9}$$

is uncorrelated with the regressors, S_i and X_i, and the regression coefficient ρ is the causal effect of interest. It bears emphasizing once again that the key assumption here is that the observable characteristics, X_i, are the only reason why η_i and S_i (equivalently, $f_i(s)$ and S_i) are correlated. This is the selection-on-observables assumption for regression models discussed over a quarter century ago by Barnow, Cain, and Goldberger (1981). It remains the basis of most empirical work in Economics.

3.2.2 The Omitted Variables Bias Formula

The omitted variables bias (OVB) formula describes the relationship between regression estimates in models with different sets of control variables. This important formula is often motivated by the notion that a longer regression, i.e., one with more controls such as equation (3.2.9), has a causal interpretation, while a shorter regression does not. The coefficients on the variables included in the shorter regression are therefore said to be "biased". In fact, the OVB formula is a mechanical link between coefficient vectors that applies to short and long regressions whether or not the longer regression is causal. Nevertheless, we follow convention and refer to the difference between the included coefficients in a long regression and a short regression as being determined by the OVB formula.

To make this discussion concrete, suppose the set of relevant control variables in the schooling regression can be boiled down to a combination of family background, intelligence and motivation. Let these specific factors be denoted by a vector, A_i, which we'll refer to by the shorthand term “ability.” The regression of

wages on schooling, S_i, controlling for ability can written as

$$\text{Y}_i = \alpha + \rho \text{S}_i + A_i'\gamma + \varepsilon_i, \tag{3.2.10}$$

where α, ρ, and γ are population regression coefficients, and ε_i is a regression residual that is uncorrelated with all regressors by definition. If the CIA applies given A_i, then ρ can be equated with the coefficient in the linear causal model, 3.2.7, while the residual ε_i is the random part of potential earnings that is left over after controlling for A_i.

In practice, ability is hard to measure. For example, the American Current Population Survey (CPS), a large data set widely used in applied microeconomics (and the source of U.S. government data on unemployment rates), tells us nothing about adult respondents' family background, intelligence, or motivation. What are the consequences of leaving ability out of regression (3.2.10)? The resulting "short regression" coefficient is related to the "long regression" coefficient in equation (3.2.10) as follows:

$$\frac{Cov(\text{Y}_i, \text{S}_i)}{V(\text{S}_i)} = \rho + \gamma'\delta_{As}, \tag{3.2.11}$$

where δ_{As} is the vector of coefficients from regressions of the elements of A_i on S_i. To paraphrase, the OVB formula says

> Short equals long plus the effect of omitted times the regression of omitted on included.

This formula is easy to derive: plug the long regression into the short regression formula, $\frac{Cov(\text{Y}_i,\text{S}_i)}{V(\text{S}_i)}$. Not surprisingly, the OVB formula is closely related to the regression anatomy formula, 3.1.3, from Section 3.1.2. Both the OVB and regression anatomy formulas tell us that short and long regression coefficients are the same whenever the omitted and included variables are uncorrelated.[10]

We can use the OVB formula to get a sense of the likely consequences of omitting ability for schooling coefficients. Ability variables have positive effects on wages, and these variables are also likely to be positively correlated with schooling. The short regression coefficient may therefore be "too big" relative to what we want. On the other hand, as a matter of economic theory, the direction of the correlation between schooling and ability is not entirely clear. Some omitted variables may be negatively correlated with schooling, in which case the short regression coefficient will be too small.[11]

[10] Here is the multivariate generalization of OVB: Let β_1^s denote the coefficient vector on a $\text{K}_1 \times 1$ vector of variables, X_{1i} in a (short) regression that has no other variables and let β_1^l denote the coefficient vector on these variables in a (long) regression that includes a $\text{K}_2 \times 1$ vector of control variables, X_{2i}, with coefficient vector β_2^l. Then $\beta_1^s = \beta_1^l + E[X_{1i}X_{1i}']^{-1}E[X_{1i}X_{2i}']\beta_2^l$.

[11] As highly educated people, we like to assume that ability and schooling are positively correlated. This is not a foregone conclusion, however: Mick Jagger dropped out of the London School of Economics and Bill Gates dropped out of Harvard, perhaps because the opportunity cost of schooling for these high-ability guys was high (of course, they may also be a couple of very lucky college dropouts).

Table 3.2.1 illustrates these points using data from the NLSY. The first three entries in the table show that the schooling coefficient decreases from .132 to .114 when family background variables—in this case, parents' education—as well as a few basic demographic characteristics (age, race, census region of residence) are included as controls. Further control for individual ability, as proxied by the Armed Forces Qualification Test (AFQT) test score, reduces the schooling coefficient to .087 (AFQT is used by the military to select soldiers). The omitted variables bias formula tells us that these reductions are a result of the fact that the additional controls are positively correlated with both wages and schooling.[12]

Table 3.2.1: Estimates of the returns to education for men in the NLSY

	(1)	(2)	(3)	(4)	(5)
Controls:	None	Age dummies	Col. (2) and additional controls*	Col. (3) and AFQT score	Col. (4), with occupation dummies
	0.132	0.131	0.114	0.087	0.066
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)

Notes: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey). The table reports the coefficient on years of schooling in a regression of log wages on years of schooling and the indicated controls. Standard errors are shown in parentheses. The sample is restricted to men and weighted by NLSY sampling weights. The sample size is 2434.

*Additional controls are mother's and father's years of schooling and dummy variables for race and Census region.

Although simple, the OVB formula is one of the most important things to know about regression. The importance of the OVB formula stems from the fact that if you claim an absence of omitted variables bias, then typically you're also saying that the regression you've got is the one you want. And the regression you want usually has a causal interpretation. In other words, you're prepared to lean on the CIA for a causal interpretation of the long-regression estimates.

At this point, it's worth considering when the CIA is most likely to give a plausible basis for empirical work. The best-case scenario is random assignment of s_i , conditional on X_i, in some sort of (possibly natural) experiment. An example is the study of a mandatory re-training program for unemployed workers by Black, *et al.* (2003). The authors of this study were interested in whether the re-training program succeeded in raising earnings later on. They exploit the fact that eligibility for the training program they study was determined on the basis of personal characteristics and past unemployment and job histories. Workers were divided up into groups on the basis of these characteristics. While some of these groups of workers were ineligible for training, those in other groups were required to take training if they did not take

[12] A large empirical literature investigates the consequences of omitting ability variables from schooling equations. Key early references include Griliches and Mason (1972), Taubman (1976), Griliches (1977), and Chamberlain (1978).

a job. When some of the mandatory training groups contained more workers than training slots, training opportunities were distributed by lottery. Hence, training requirements were randomly assigned conditional on the covariates used to assign workers to groups. A regression on a dummy for training plus the personal characteristics, past unemployment variables, and job history variables used to classify workers seems very likely to provide reliable estimates of the causal effect of training.[13]

In the schooling context, there is usually no lottery that directly determines whether someone will go to college or finish high school.[14] Still, we might imagine subjecting individuals of similar ability and from similar family backgrounds to an experiment that encourages school attendance. The Education Maintenance Allowance, which pays British high school students in certain areas to attend school, is one such policy experiment (Dearden, et al, 2004).

A second type of study that favors the CIA exploits detailed institutional knowledge regarding the process that determines s_i . An example is the Angrist (1998) study of the effect of voluntary military service on the later earnings of soldiers. This research asks whether men who volunteered for service in the US Armed Forces were economically better off in the long run. Since voluntary military service is not randomly assigned, we can never know for sure. Angrist therefore used matching and regression techniques to control for observed differences between veterans and nonveterans who applied to get into the all-volunteer forces between 1979 and 1982. The motivation for a control strategy in this case is the fact that the military screens soldier-applicants primarily on the basis of observable covariates like age, schooling, and test scores.

The CIA in Angrist (1998) amounts to the claim that after conditioning on all these observed characteristics veterans and nonveterans are comparable. This assumption seems worth entertaining since, conditional on X_i, variation in veteran status in the Angrist (1998) study comes solely from the fact that some qualified applicants fail to enlist at the last minute. Of course, the considerations that lead a qualified applicant to "drop out" of the enlistment process could be related to earnings potential, so the CIA is clearly not guaranteed even in this case.

3.2.3 Bad Control

We've made the point that control for covariates can make the CIA more plausible. But more control is not always better. Some variables are bad controls and should not be included in a regression model even when their inclusion might be expected to change the short regression coefficients. Bad controls are variables that are themselves outcome variables in the notional experiment at hand. That is, bad controls might just as well be dependent variables too. Good controls are variables that we can think of as having been fixed at the time the regressor of interest was determined.

The essence of the bad control problem is a version of selection bias, albeit somewhat more subtle than

[13] This program appears to raise earnings, primarily because workers in the training group went back to work more quickly.

[14] Lotteries have been used to distribute private school tuition subsidies; see, e.g., Angrist, et al. (2002).

the selection bias discussed in Chapter (2) and Section (3.2). To illustrate, suppose we are interested in the effects of a college degree on earnings and that people can work in one of two occupations, white collar and blue collar. A college degree clearly opens the door to higher-paying white collar jobs. Should occupation therefore be seen as an omitted variable in a regression of wages on schooling? After all, occupation is highly correlated with both education and pay. Perhaps it's best to look at the effect of college on wages for those within an occupation, say white collar only. The problem with this argument is that once we acknowledge the fact that college affects occupation, comparisons of wages by college degree status within an occupation are no longer apples-to-apples, *even if college degree completion is randomly assigned.*

Here is a formal illustration of the bad control problem in the college/occupation example.[15] Let W_i be a dummy variable that denotes white collar workers and let Y_i denote earnings. The realization of these variables is determined by college graduation status and potential outcomes that are indexed against C_i. We have

$$\begin{aligned} \mathrm{Y}_i &= \mathrm{C}_i\mathrm{Y}_{1i} + (1-\mathrm{C}_i)\mathrm{Y}_{0i} \\ \mathrm{W}_i &= \mathrm{C}_i\mathrm{W}_{1i} + (1-\mathrm{C}_i)\mathrm{W}_{0i} \end{aligned}$$

where $\mathrm{C}_i = 1$ for college graduates and is zero otherwise, $\{\mathrm{Y}_{1i},\mathrm{Y}_{0i}\}$ denotes potential earnings, and $\{\mathrm{W}_{1i},\mathrm{W}_{0i}\}$ denotes potential white-collar status. We assume that C_i is randomly assigned, so it is independent of all potential outcomes. We have no trouble estimating the causal effect of C_i on either Y_i or W_i since independence gives us

$$\begin{aligned} E\left[\mathrm{Y}_i|\mathrm{C}_i=1\right] - E\left[\mathrm{Y}_i|\mathrm{C}_i=0\right] &= E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}\right], \\ E\left[\mathrm{W}_i|\mathrm{C}_i=1\right] - E\left[\mathrm{W}_i|\mathrm{C}_i=0\right] &= E\left[\mathrm{W}_{1i}-\mathrm{W}_{0i}\right]. \end{aligned}$$

In practice, we might estimate these average treatment effects by regressing Y_i and W_i and on C_i.

Bad control means that a comparison of earnings conditional on W_i does not have a causal interpretation. Consider the difference in mean earnings between college graduates and others conditional on working at a white collar job. We can compute this in a regression model that includes W_i or by regressing Y_i on C_i in the sample where $\mathrm{W}_i = 1$. The estimand in the latter case is the difference in means with C_i switched off and on, conditional on $\mathrm{W}_i = 1$:

$$E\left[\mathrm{Y}_i|\mathrm{W}_i=1,\mathrm{C}_i=1\right] - E\left[\mathrm{Y}_i|\mathrm{W}_i=1,\mathrm{C}_i=0\right] = E\left[\mathrm{Y}_{1i}|\mathrm{W}_{1i}=1,\mathrm{C}_i=1\right] - E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1,\mathrm{C}_i=0\right] \quad (3.2.12)$$

[15] The same problem arises in "conditional-on-positive" comparisons, discussed in detail in section (3.4.2), below.

By the joint independence of $\{\mathrm{Y}_{1i},\mathrm{W}_{1i},\mathrm{Y}_{0i},\mathrm{W}_{0i}\}$ and C_i, we have

$$E\left[\mathrm{Y}_{1i}|\mathrm{W}_{1i}=1,\mathrm{C}_i=1\right]-E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1,\mathrm{C}_i=0\right]=E\left[\mathrm{Y}_{1i}|\mathrm{W}_{1i}=1\right]-E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1\right].$$

This expression illustrates the apples-to-oranges nature of the bad-control problem:

$$\begin{aligned}&E\left[\mathrm{Y}_{1i}|\mathrm{W}_{1i}=1\right]-E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1\right]\\&=\underbrace{E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{W}_{1i}=1\right]}_{\text{causal effect on college grads}}+\underbrace{\left\{E\left[\mathrm{Y}_{0i}|\mathrm{W}_{1i}=1\right]-E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1\right]\right\}}_{\text{selection bias}}.\end{aligned}$$

In other words, the difference in wages between those with and without a college degree conditional on working in a white collar job equals the causal effect of college on those with $\mathrm{W}_{1i}=1$ (people who work at a white collar job when they have a college degree) and a selection-bias term which reflects the fact that college changes the composition of the pool of white collar workers.

The selection-bias in this context can be positive or negative, depending on the relation between occupational choice, college attendance, and potential earnings. The main point is that even if $\mathrm{Y}_{1i}=\mathrm{Y}_{0i}$, so that there is no causal effect of college on wages, the conditional comparison in (3.2.12) will not tell us this (the regression of Y_i on W_i and C_i has exactly the same problem). It is also incorrect to say that the conditional comparison captures the part of the effect of college that is "not explained by occupation." In fact, the conditional comparison does not tell us much that is useful without a more elaborate model of the links between college, occupation, and earnings.[16]

As an empirical illustration, we see that the addition of two-digit occupation dummies indeed reduces the schooling coefficient in the NLSY models reported in Table 3.2.1, in this case from .087 to .066. However, it's hard to say what we should make of this decline. The change in schooling coefficients when we add occupation dummies may simply be an artifact of selection bias. So we would do better to control only for variables that are not themselves caused by education.

A second version of the bad control scenario involves *proxy control*, that is, the inclusion of variables that might partially control for omitted factors, but are themselves affected by the variable of interest. A simple version of the proxy-control scenario goes like this: Suppose you are interested in a long regression, similar to equation (3.2.10),

$$\mathrm{Y}_i=\alpha+\rho\mathrm{S}_i+\gamma a_i+\varepsilon_i, \tag{3.2.13}$$

where for the purposes of this discussion we've replaced the vector of controls A_i, with a scalar ability measure a_i. Think of this as an IQ score that measures innate ability in eighth grade, before any relevant

[16] In this example, selection bias is probably negative, that is $E\left[\mathrm{Y}_{0i}|\mathrm{W}_{1i}=1\right]<E\left[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1\right]$. It seems reasonable to think that any college graduate can get a white collar job, so $E\left[\mathrm{Y}_{0i}|\mathrm{W}_{1i}=1\right]$ is not too far from $E[\mathrm{Y}_{0i}]$. But someone who gets a white collar without benefit of a college degree (i.e., $\mathrm{W}_{0i}=1$) is probably special, i.e., has a better than average Y_{0i}.

schooling choices are made (assuming everyone completes eighth grade). The error term in this equation satisfies $E[\text{S}_i\varepsilon_i] = E[a_i\varepsilon_i] = 0$ by definition. Since a_i is measured before S_i is determined, it is a good control.

Equation (3.2.13) is the regression of interest, but unfortunately, data on a_i are unavailable. However, you have a second ability measure collected later, after schooling is completed (say, the score on a test used to screen job applicants). Call this variable "late ability," a_{li}. In general, schooling increases late ability relative to innate ability. To be specific, suppose

$$a_{li} = \pi_0 + \pi_1\text{S}_i + \pi_2 a_i. \tag{3.2.14}$$

By this, we mean to say that both schooling and innate ability increase late or measured ability. There is almost certainly some randomness in measured ability as well, but we can make our point more simply via the deterministic link, (3.2.14).

You're worried about OVB in the regression of Y_i on S_i alone, so you propose to regress Y_i on S_i and late ability, a_{li} since the desired control, a_i, is unavailable. Using (3.2.14) to substitute for a_i in (3.2.13), the regression on S_i and a_{li} is

$$\text{Y}_i = (\alpha - \gamma\frac{\pi_0}{\pi_2}) + (\rho - \gamma\frac{\pi_1}{\pi_2})\text{S}_i + \frac{\gamma}{\pi_2}a_{li} + \varepsilon_i. \tag{3.2.15}$$

In this scenario, γ, π_1, and π_2 are all positive, so $\rho - \gamma\frac{\pi_1}{\pi_2}$ is too small unless π_1 turns out to be zero. In other words, use of a proxy control that is increased by the variable of interest generates a coefficient below the desired effect. Importantly, π_1 can be investigated to some extent: if the regression of a_{li} on S_i is zero, you might feel better about assuming that π_1 is zero in (3.2.14).

There is an interesting ambiguity in the proxy-control story that is not present in the first bad-control story. Control for outcome variables is simply misguided; you do not want to control for occupation in a schooling regression if the regression is to have a causal interpretation. In the proxy-control scenario, however, your intentions are good. And while proxy control does not generate the regression coefficient of interest, it may be an improvement on no control at all. Recall that the motivation for proxy control is equation (3.2.13). In terms of the parameters in this model, the OVB formula tells us that a regression on S_i with no controls generates a coefficient of $\rho + \gamma\delta_{as}$, where δ_{as} is slope coefficient from a regression of a_i on S_i. The schooling coefficient in (3.2.15) might be closer to ρ than the coefficient you estimate with no control at all. Moreover, assuming δ_{as} is positive, you can safely say that the causal effect of interest lies between these two.

One moral of both the bad-control and the proxy-control stories is that when thinking about controls, timing matters. Variables measured before the variable of interest was determined are generally good controls. In particular, because these variables were determined before the variable of interest, they cannot themselves

be outcomes in the causal nexus. In many cases, however, the timing is uncertain or unknown. In such cases, clear reasoning about causal channels requires explicit assumptions about what happened first, or the assertion that none of the control variables are themselves caused by the regressor of interest.[17]

3.3 Heterogeneity and Nonlinearity

As we saw in the previous section, a linear causal model in combination with the CIA leads to a linear CEF with a causal interpretation. Assuming the CEF is linear, the population regression is it. In practice, however, the assumption of a linear CEF is not really necessary for a causal interpretation of regression. For one thing, as discussed in Section 3.1.2, we can think of the regression of Y_i on X_i and S_i as providing the best linear approximation to the underlying CEF, regardless of its shape. Therefore, if the CEF is causal, the fact that regression approximates it gives regression coefficients a causal flavor. This claim is a little vague, however, and the nature of the link between regression and the CEF is worth exploring further. This exploration leads us to an understanding of regression as a computationally attractive matching estimator.

3.3.1 Regression Meets Matching

The past decade or two has seen increasing interest in matching as an empirical tool. Matching as a strategy to control for covariates is typically motivated by the CIA, as for causal regression in the previous section. For example, Angrist (1998) used matching to estimate the effects of volunteering for the military service on the later earnings of soldiers. These matching estimates have a causal interpretation assuming that, conditional on the individual characteristics the military uses to select soldiers (age, schooling, test scores), veteran status is independent of potential earnings.

An attractive feature of matching strategies is that they are typically accompanied by an explicit statement of the conditional independence assumption required to give matching estimates a causal interpretation. At the same time, we have just seen that the causal interpretation of a regression coefficient is based on exactly the same assumption. In other words, matching and regression are both control strategies. Since the core assumption underlying causal inference is the same for the two strategies, it's worth asking whether or to what extent matching really differs from regression. Our view is that regression can be motivated as a computational device for a particular sort of weighted matching estimator, and therefore the differences between regression and matching are unlikely to be of major empirical importance.

To flesh out this idea, it helps to look more deeply into the mathematical structure of the matching and regressions *estimands*, i.e., the population quantities that these methods attempt to estimate. For regression, of course, the estimand is a vector of population regression coefficients. The matching estimand is typically

[17] Griliches and Mason (1972) is a seminal exploration of the use of early and late ability controls in schooling equations. See also Chamberlain (1977, 1978) for closely related studies. Rosenbaum (1984) offers an alternative discussion of the proxy control idea using very different notation, outside of a regression framework.

a particular weighted average of contrasts or comparisons across cells defined by covariates. This is easiest to see in the case of discrete covariates, as in the military service example, and for a discrete regressor such as veteran status, which we denote here by the dummy, D_i. Since treatment takes on only two values, we can use the notation $Y_{1i}=f_i(1)$ and $Y_{0i}=f_i(0)$ to denote potential outcomes. A parameter of primary interest in this context is the average effect of treatment on the treated, $E[Y_{1i}-Y_{0i}|D_i = 1]$. This tells us the difference between the average earnings of soldiers, $E[Y_{1i}|D_i = 1]$, an observable quantity, and the counterfactual average earnings they would have obtained if they had not served, $E[Y_{0i}|D_i = 1]$. Simply comparing the observed earnings differential by veteran status is a biased measure of the effect of treatment on the treated unless D_i is independent of Y_{0i}. Specifically,

$$\begin{aligned} E\left[Y_i|D_i = 1\right] - E\left[Y_i|D_i = 0\right] &= E\left[Y_{1i} - Y_{0i}|D_i = 1\right] \\ &\quad + \left\{E\left[Y_{0i}|D_i = 1\right] - E\left[Y_{0i}|D_i = 0\right]\right\}. \end{aligned}$$

In other words, the observed earnings difference by veteran status equals the average effect of treatment on the treated plus selection bias. This parallels the discussion of selection bias in Chapter 2.

Given the CIA, selection bias disappears after conditioning on X_i, so the effect of treatment on the treated can be constructed by iterating expectations over X_i:

$$\begin{aligned} \delta_{TOT} &\equiv E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E\{E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 1]|D_i = 1\}. \end{aligned}$$

Of course, $E[Y_{0i}|X_i,D_i = 1]$ is counterfactual. By virtue of the CIA, however,

$$E[Y_{0i}|X_i, D_i = 0] = E[Y_{0i}|X_i, D_i = 1].$$

Therefore,

$$\begin{aligned} \delta_{TOT} &= E\left\{E\left[Y_{1i}|X_i, D_i = 1\right] - E\left[Y_{0i}|X_i, D_i = 0\right]|D_i = 1\right\} \\ &= E[\delta_X|D_i = 1], \end{aligned} \tag{3.3.1}$$

where

$$\delta_X \equiv E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0],$$

is the random X-specific difference in mean earnings by veteran status at each value of X_i.

The matching estimator in Angrist (1998) uses the fact that X_i is discrete to construct the sample analog

of the right-hand-side of (3.3.1). In the discrete case, the matching estimand can be written

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1] = \sum_x \delta_x P(\mathrm{X}_i = x|\mathrm{D}_i = 1), \tag{3.3.2}$$

where $P(\mathrm{X}_i = x|\mathrm{D}_i = 1)$ is the probability mass function for X_i given $\mathrm{D}_i = 1$.[18]. In this case, X_i, takes on values determined by all possible combinations of year of birth, test-score group, year of application to the military, and educational attainment at the time of application. The test score in this case is from the AFQT, used by the military to categorize the mental abilities of applicants (we included this as a control in the schooling regression discussed in Section 3.2.2). The Angrist (1998) matching estimator simply replaces δ_X by the sample veteran-nonveteran earnings difference for each combination of covariates, and then combines these in a weighted average using the empirical distribution of covariates among veterans.[19]

Note also that we can just as easily construct the unconditional average treatment effect,

$$\begin{aligned} \delta_{ATE} &= E\{E[\mathrm{Y}_{1i}|\mathrm{X}_i, \mathrm{D}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{X}_i, \mathrm{D}_i = 0]\} \\ &= \sum_x \delta_x P(\mathrm{X}_i = x) \\ &= E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}], \end{aligned} \tag{3.3.3}$$

which is the expectation of δ_X using the marginal distribution of X_i instead of the distribution among the treated. δ_{TOT} tells us how much the typical soldier gained or lost as a consequence of military service, while δ_{ATE} tells us how much the typical applicant to the military gained or lost (since the Angrist, 1998, population consists of applicants.)

The US military tends to be fairly picky about it's soldiers, especially after downsizing at the end of the Cold War. For the most part, the military now takes only high school graduates with test scores in the upper half of the test score distribution. The resulting positive screening generates positive selection bias in naive comparisons of veteran and non-veteran earnings. This can be seen in Table 3.3.1, which reports differences-in-means, matching, and regression estimates of the effect voluntary military service on the 1988-91 Social Security-taxable earnings of men who applied to join the military between 1979 and 1982. The matching estimates were constructed from the sample analog of (3.3.2). Although white veterans earn \$1,233 more than nonveterans, this difference becomes negative once differences in covariates are matched away. Similarly, while non-white veterans earn \$2,449 more than nonveterans, controlling for covariates reduces this to \$840.

[18] This matching estimator is discussed by Rubin (1977) and used by Card and Sullivan (1988) to estimate the effect of subsidized training on employment.

[19] With continuous covariates, exact matching is impossible and some sort of approximation is required, a fact that leads to bias. See Abadie and Imbens (2006), who derive the implications of approximate matching for the limiting distirbution of matching estimators.

Table 3.3.1: Uncontrolled, matching, and regression estimates of the effects of voluntary military service on earnings

Race	Average earnings in 1988-1991	Differences in means by veteran status	Matching estimates	Regression estimates	Regression minus matching
	(1)	(2)	(3)	(4)	(5)
Whites	14537	1233.4	-197.2	-88.8	108.4
		(60.3)	(70.5)	(62.5)	(28.5)
Non-whites	11664	2449.1	839.7	1074.4	234.7
		(47.4)	(62.7)	(50.7)	(32.5)

Notes: Adapted from Angrist (1998, Tables II and V). Standard errors are reported in parentheses. The table shows estimates of the effect of voluntary military service on the 1988-1991 Social Security- taxable earnings of men who applied to enter the armed forces between 1979 and 1982. The matching and regression estimates control for applicants' year of birth, education at the time of application, and AFQT score. There are 128,968 whites and 175,262 non-whites in the sample.

Table (3.3.1) also shows regression estimates of the effect of voluntary military service, controlling for the same set of covariates that were used to construct the matching estimates. These are estimates of δ_R in the equation

$$\mathrm{Y}_i = \sum_x d_{ix}\beta_x + \delta_R \mathrm{D}_i + \varepsilon_i, \tag{3.3.4}$$

where d_{ix} is a dummy that indicates $\mathrm{X}_i = x$, β_x is a regression-effect for $\mathrm{X}_i = x$, and δ_R is the regression estimand. Note that this regression model allows a separate parameter for every value taken on by the covariates. This model can therefore be said to be saturated-in-X_i, since it includes a parameter for every value of X_i (it is not "fully saturated," however, because there is a single additive effect for D_i with no $\mathrm{D}_i \cdot \mathrm{X}_i$ interactions).

Despite the fact that the matching and regression estimates control for the same variables, the regression estimates in Table 3.3.1 are somewhat larger than the matching estimates for both whites and nonwhites. In fact, the differences between the matching and regression results are statistically significant. At the same time, the two estimation strategies present a broadly similar picture of the effects of military service. The reason the regression and matching estimates are similar is that regression, too, can be seen as a sort of matching estimator: the regression estimand differs from the matching estimands only in the weights used to sum the covariate-specific effects, δ_X into a single effect. In particular, matching uses the distribution of covariates among the treated to weight covariate-specific estimates into an estimate of the effect of treatment on the treated, while regression produces a variance-weighted average of these effects.

To see this, start by using the regression anatomy formula to write the coefficient on D_i in the regression of Y_i on X_i and D_i as

$$\begin{aligned} \delta_R &= \frac{Cov(\mathrm{Y}_i, \tilde{\mathrm{D}}_i)}{V(\tilde{\mathrm{D}}_i)} && (3.3.5) \\ &= \frac{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])\mathrm{Y}_i]}{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2]} \\ &= \frac{E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])E[\mathrm{Y}_i|\mathrm{D}_i, \mathrm{X}_i]\}}{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2]}. && (3.3.6) \end{aligned}$$

The second equality in this set of expressions uses the fact that saturating the model in X_i means $E[\mathrm{D}_i|\mathrm{X}_i]$ is linear. Hence, $\tilde{\mathrm{D}}_i$, which is defined as the residual from a regression of D_i on X_i, is the difference between D_i and $E[\mathrm{D}_i|\mathrm{X}_i]$. The third equality uses the fact that the regression of Y_i on D_i and X_i is the same as the regression of Y_i on $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i]$.

To simplify further, we expand the CEF, $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i]$, to get

$$E[\mathrm{Y}_i|\mathrm{D}_i, \mathrm{X}_i] = E[\mathrm{Y}_i|\mathrm{D}_i = 0, \mathrm{X}_i] + \delta_X \mathrm{D}_i.$$

If covariates are unnecessary - in other words, the CIA holds unconditionally, as if in a randomized trial - this CEF becomes

$$E[\mathrm{Y}_i|\mathrm{D}_i, \mathrm{X}_i] = E[\mathrm{Y}_i|\mathrm{D}_i = 0] + E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}]\mathrm{D}_i,$$

from which we conclude that the regression of Y_i on D_i estimates the population average treatment effect in this case (e.g., as in the experiment discussed in Section 2.3). But here we are interested in the more general scenario where conditioning X_i is necessary to eliminate selection bias.

To evaluate the more general regression estimand, (3.3.5), we begin by substituting for $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i]$ in the numerator. This gives

$$E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])E[\mathrm{Y}_i|\mathrm{D}_i, \mathrm{X}_i]\} = E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])E[\mathrm{Y}_i|\mathrm{D}_i = 0, \mathrm{X}_i]\} + E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])\mathrm{D}_i\delta_X\}.$$

The first term on the right-hand side is zero because $E[\mathrm{Y}_i|\mathrm{D}_i = 0,\mathrm{X}_i]$ is a function of X_i and is therefore uncorrelated with $(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])$. For the same reason, the second term simplifies to

$$E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])\mathrm{D}_i\delta_X\} = E\{(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2\delta_X\}.$$

At this point, we've shown

$$\delta_R = \frac{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2\delta_X]}{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2]} = \frac{E\{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]\delta_X\}}{E\{E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]\}} = \frac{E[\sigma^2_D(\mathrm{X}_i)\delta_X]}{E[\sigma^2_D(\mathrm{X}_i)]}, \quad (3.3.7)$$

where

$$\sigma_D^2(\mathrm{X}_i) = E[(\mathrm{D}_i - E[\mathrm{D}_i|\mathrm{X}_i])^2|\mathrm{X}_i]$$

is the conditional variance of D_i given X_i. This establishes that the regression model, (3.3.4), produces a treatment-variance weighted average of δ_X.

Because the regressor of interest, D_i is a dummy variable, one last step can be taken. In this case, $\sigma_D^2(\mathrm{X}_i) = P(\mathrm{D}_i = 1|\mathrm{X}_i)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i))$, so

$$\delta_R = \frac{\sum_x \delta_x \left[P(\mathrm{D}_i = 1|\mathrm{X}_i = x)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i = x))\right] P\left(\mathrm{X}_i = x\right)}{\sum_x \left[P(\mathrm{D}_i = 1|\mathrm{X}_i = x)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i = x))\right] P\left(\mathrm{X}_i = x\right)}$$

This shows that the regression estimand weights each covariate-specific treatment effect by $[P(\mathrm{X}_i = x|\mathrm{D}_i = 1)(1 - P(\mathrm{X}_i = x|\mathrm{D}_i = 1))]P\left(\mathrm{X}_i = x\right)$. In contrast, the matching estimand for the effect of treatment on the treated can be written

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1] = \sum_x \delta_x P(\mathrm{X}_i = x|\mathrm{D}_i = 1) = \frac{\sum_x \delta_x P(\mathrm{D}_i = 1|\mathrm{X}_i = x)P(\mathrm{X}_i = x)}{\sum_x P(\mathrm{D}_i = 1|\mathrm{X}_i = x)P(\mathrm{X}_i = x)}$$

because

$$P(\mathrm{X}_i = x|\mathrm{D}_i = 1) = \frac{P(\mathrm{D}_i = 1|\mathrm{X}_i = x) \cdot P(\mathrm{X}_i = x)}{P(\mathrm{D}_i = 1)}.$$

So the weights used to construct $E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1]$ are proportional to the probability of treatment at each value of the covariates.

The point of this derivation is that the treatment-on-the-treated estimand puts the most weight on covariate cells containing those who are most likely to be treated. In contrast, regression puts the most weight on covariate cells where the conditional variance of treatment status is largest. As a rule, this variance is maximized when $P(\mathrm{D}_i = 1|\mathrm{X}_i = x) = \frac{1}{2}$, in other words, for cells where there are equal numbers of treated and control observations. Of course, the difference in weighting schemes is of little importance if δ_x does not vary across cells (though weighting still affects the statistical efficiency of estimators). In this example, however, men who were most likely to serve in the military appear to benefit least from their service. This is probably because those most likely to serve were most qualified, but therefore also had the highest civilian earnings potential and so benefited least from military service. This fact leads matching estimates of the effect of military service to be smaller than regression estimates based on the same vector of control variables.[20]

[20] It's no surprise that regression gives the most weight to cells where $P(\mathrm{D}_i = 1|\mathrm{X}_i = x) = 1/2$ since regression is efficient for a homoskedastic constant-effects linear model. We should expect an efficient estimator to give the most weight to cells where the common treatment effect is estimated most precisely. With homoskedastic residuals, the most precise treatment effects

Importantly, neither the regression nor the covariate-matching estimands give any weight to covariate cells that do not contain both treated and control observations. Consider a value of X_i, say x^*, where either no one is treated or everyone is treated. Then, δ_{x^*} is undefined, while the regression weights, $[P(\mathrm{D}_i = 1|\mathrm{X}_i = x^*)(1 - P(\mathrm{D}_i = 1|\mathrm{X}_i = x^*))]$, are zero. In the language of the econometric literature on matching, both the regression and matching estimands impose *common support*, that is, they are limited to covariate values where both treated and control observations are found.[21]

The step from *estimand* to *estimator* is a little more complicated. In practice, both regression and matching estimators are implemented using modelling assumptions that implicitly involve a certain amount of extrapolation across cells. For example, matching estimators often combine covariates cells with few observations. This violates common support if the cells being combined do not each have both treated and non-treated observations. Regression models that are not saturated in X_i may also violate common support, since covariate cells without both treated and control observations can end up contributing to the estimates by extrapolation. Here too, however, we see a symmetry between the matching and regression strategies: they are in the same class, in principle, and require the same sort of compromises in practice.[22]

Even More on Regression and Matching: Ordered and Continuous Treatments★

Does the pseudo-matching interpretation of regression outlined above for a binary treatment apply to models with ordered and continuous treatments? The long answer is fairly technical and may be more than you want to know. The short answer is, to one degree or another, "yes."

As we've already discussed, one interpretation of regression is that the population OLS slope vector provides the MMSE linear approximation to the CEF. This, of course, works for ordered and continuous regressors as well as for binary. A related property is the fact that regression coefficients have an "average derivative" interpretation. In multivariate regression models, this interpretation is unfortunately complicated by the fact that the OLS slope vector is a matrix-weighted average of the gradient of the CEF. Matrix-weighted averages are difficult to interpret except in special cases (see Chamberlain and Leamer, 1976). An important special case when the average derivative property is relatively straightforward is in regression models for an ordered or continuous treatment with a saturated model for covariates. To avoid lengthy derivations, we simply explain the formulas. A derivation is sketched in the appendix to this chapter. For additional details, see the appendix to Angrist and Krueger (1999).

---

come from cells where the probability of treatment equals 1/2.

[21] The *support* of a random variable is the set of realizations that occur with positive probability. See Heckman, Ichimura, Smith, and Todd (1998) and Smith and Todd (2001) for a discussion of common support in matching.

[22] Matching problems involving finely distributed X-variables are often solved by aggregating values to make coarser groupings or by pairing observations that have similar, though not necessarily identical values. See Cochran (1965), Rubin (1973), or Rosenbaum (1995, Chapter 3) for discussions of this approach. With continuously-distributed covariates, matching estimators are biased because matches are imperfect. Abadie and Imbens (2008) have recently shown that a regression-based bias correction can eliminate the (asymptotic) bias from imperfect matches.

For the purposes of this discussion, the treatment intensity, S_i, is assumed to be a continuously distributed random variable, not necessarily non-negative. Suppose that the CEF of interest can be written $h(t) \equiv E[\mathrm{Y}_i|\mathrm{S}_i = t]$ with derivative $h'(t)$. Then

$$\frac{E[\mathrm{Y}_i(\mathrm{S}_i - E[\mathrm{S}_i])]}{E[\mathrm{S}_i(\mathrm{S}_i - E[\mathrm{S}_i])]} = \frac{\int h'(t)\,\mu_t dt}{\int \mu_t dt} \tag{3.3.8}$$

where

$$\mu_t \equiv \{E[\mathrm{S}_i|\mathrm{S}_i \geq t] - E[\mathrm{S}_i|\mathrm{S}_i < t]\}\{P(\mathrm{S}_i \geq t)[1 - P(\mathrm{S}_i \geq t)\}, \tag{3.3.9}$$

and the integrals in (3.3.8) run over the possible values of S_i. This formula weights each possible value of S_i in proportion to the difference in the conditional mean of S_i above and below that value. More weight is also given to points close to the median of S_i since $P(\mathrm{S}_i \geq t) \cdot [1 - P(\mathrm{S}_i \geq t)]$ is maximized at $P(\mathrm{S}_i \geq t) = 1/2$.

With covariates, X_i, the weights in (3.3.8) become X-specific. A covariate-averaged version of the same formula applies to the multivariate regression coefficient of Y_i on S_i, after partialling out X_i. In particular,

$$\frac{E[\mathrm{Y}_i(\mathrm{S}_i - E[\mathrm{S}_i|\mathrm{X}_i])]}{E[\mathrm{S}_i(\mathrm{S}_i - E[\mathrm{S}_i|X])]} = \frac{E\left[\int h'_X(t)\mu_{tX} dt\right]}{E\left[\int \mu_{tX} dt\right]}, \tag{3.3.10}$$

where $h'_X(t) \equiv \frac{\partial E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{S}_i = t]}{\partial t}$ and $\mu_{tX} \equiv \{E[\mathrm{S}_i|\mathrm{X}_i, \mathrm{S}_i \geq t] - E[\mathrm{S}_i|\mathrm{X}_i, \mathrm{S}_i < t]\}\{P(\mathrm{S}_i \geq t|\mathrm{X}_i)[1 - P(\mathrm{S}_i \geq t|\mathrm{X}_i)\}$. It bears emphasizing that equation (3.3.10) reflects two types of averaging: an integral that averages *along* the length of a nonlinear CEF at fixed covariate values, and an expectation that averages *across* covariate cells. An important point in this context is that population regression coefficients contain no information about the effect of S_i on the CEF for values of X_i where $P(\mathrm{S}_i \geq t|\mathrm{X}_i)$ equals 0 or 1. This includes values of X_i where S_i is fixed. In the same spirit, it's worth noting that if S_i is a dummy variable, we can extract equation (3.3.7) from the more general formula, (3.3.10).

Angrist and Krueger (1999) construct the average weighting function for a schooling regression with state of birth and year of birth covariates. Although equations (3.3.8) and (3.3.10) may seem arcane or at least non-obvious, in this example the average weights, $E[\mu_{tX}]$, turn out to be a reasonably smooth symmetric function of t, centered at the mode of S_i.

The implications of (3.3.8) or (3.3.10) can be explored further given a model for the distribution of regressors. Suppose, for example, that S_i is Normally distributed. Let $z_i = \frac{\mathrm{S}_i - E(\mathrm{S}_i)}{\sigma_s}$, where σ_s is the standard deviation of S_i, so that z_i is standard Normal. Then

$$E[\mathrm{S}_i|\mathrm{S}_i \geq t] = E(\mathrm{S}_i) + \sigma_s E\left[z_i|z_i \geq \frac{\mathrm{t} - E(\mathrm{S}_i)}{\sigma_s}\right] = E(\mathrm{S}_i) + \sigma_s E\left[z_i|z_i \geq t^*\right].$$

From truncated Normal formulas (see, e.g., Johnson and Kotz, 1970), we know that

$$E[z_i|z_i > t^*] = \frac{\phi(t^*)}{[1-\Phi(t^*)]} \text{and } E[z_i|z_i < t^*] = \frac{-\phi(t^*)}{\Phi(t^*)}.$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard Normal density and distribution function. Substituting in the formula for μ_t, (3.3.9), we have

$$\mu_t = \sigma_s \left\{ \frac{\phi(t^*)}{[1-\Phi(t^*)]} - \frac{-\phi(t^*)}{\Phi(t^*)} \right\} [1-\Phi(t^*)]\Phi(t^*) = \sigma_s \phi(t^*).$$

We have therefore shown that

$$\frac{Cov(\mathrm{Y}_i, \mathrm{S}_i)}{V(\mathrm{S}_i)} = E[h'(\mathrm{S}_i)].$$

In other words, the regression of Y_i on S_i is the (unweighted!) population average derivative, $E[h'(\mathrm{S}_i)]$, when S_i is Normally distributed. Of course, this result is a special case of a special case.[23] Still, it seems reasonable to imagine that Normality might not matter very much. And in our empirical experience, the average derivatives (also called "marginal effects") constructed from parametric nonlinear models for limited dependent variables (e.g., Probit or Tobit) are usually indistinguishable from the corresponding regression coefficients, regardless of the distribution of regressors. We expand on this point in Section 3.4.2, below.

3.3.2 Control for Covariates Using the Propensity Score

The most important result in regression theory is the omitted variables bias formula: coefficients on included variables are unaffected by the omission of variables when the variables omitted are uncorrelated with the variables included. The propensity score theorem, due to Rosenbaum and Rubin (1983), extends this idea to estimation strategies that rely on matching instead of regression, where the causal variable of interest is a treatment dummy.[24]

The propensity score theorem states that if potential outcomes are independent of treatment status conditional on a multivariate covariate vector, X_i, then potential outcomes are independent of treatment status conditional on a scalar function of covariates, the propensity score, defined as $p(\mathrm{X}_i) \equiv E[\mathrm{D}_i|\mathrm{X}_i]$. Formally, we have

Theorem 3.3.1 *The Propensity-Score Theorem.*

Suppose the CIA holds for Y_{ji}*;* $j = 0, 1$*. Then* $\mathrm{Y}_{ji} \amalg \mathrm{D}_i | p(\mathrm{X}_i)$*.*

[23] More specialized results in this spirit appear in Ruud (1986), who considers distribution-free estimation of limited-dependent-variable models with Normally distributed regressors.

[24] Propensity-score methods can be adapted to multi-valued treatments, though this has yet to catch on. See Imbens (2000) for an effort in this direction.

Proof. The claim is true if $P[\mathrm{D}_i = 1|\mathrm{Y}_{ji}, p(\mathrm{X}_i)]$ does not depend on Y_{ji}.

$$\begin{aligned} P[\mathrm{D}_i = 1|\mathrm{Y}_{ji}, p(\mathrm{X}_i)] &= E[\mathrm{D}_i|\mathrm{Y}_{ji}, p(\mathrm{X}_i)] \\ &= E\{E[\mathrm{D}_i|\mathrm{Y}_{ji}, p(\mathrm{X}_i), \mathrm{X}_i]|\mathrm{Y}_{ji}, p(\mathrm{X}_i)\} \\ &= E\{E[\mathrm{D}_i|\mathrm{Y}_{ji}, \mathrm{X}_i]|\mathrm{Y}_{ji}, p(\mathrm{X}_i)\} \\ &= E\{E[\mathrm{D}_i|\mathrm{X}_i]|\mathrm{Y}_{ji}, p(\mathrm{X}_i)\}, \text{by the CIA.} \end{aligned}$$

But $E\{E[\mathrm{D}_i|\mathrm{X}_i]|\mathrm{Y}_{ji}, p(\mathrm{X}_i)\} = E\{p(\mathrm{X}_i)|\mathrm{Y}_{ji}, p(\mathrm{X}_i)\}$, which is clearly just $p(\mathrm{X}_i)$. ■

Like the OVB formula for regression, the propensity score theorem says you need only control for covariates that affect the probability of treatment. But it also says something more: the only covariate you really need to control for is the probability of treatment itself. In practice, the propensity score theorem is usually used for estimation in two steps: first, $p(\mathrm{X}_i)$ is estimated using some kind of parametric model, say, Logit or Probit. Then estimates of the effect of treatment are computed either by matching on the fitted values from this first step, or by a weighting scheme described below (see, Imbens, 2004, for an overview).

In practice there are many ways to use the propensity score theorem for estimation. Direct propensity-score matching works like covariate matching, except that we match on the score instead of the covariates directly. By the propensity score theorem and the CIA,

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1] = E\left\{E[\mathrm{Y}_i|p(\mathrm{X}_i), \mathrm{D}_i = 1] - E[\mathrm{Y}_i|p(\mathrm{X}_i), \mathrm{D}_i = 0]|\mathrm{D}_i = 1\right\}.$$

Estimates of the effect of treatment on the treated can therefore be obtained by stratifying on an estimate of $p(\mathrm{X}_i)$ and substituting conditional sample averages for expectations or by matching each treated observation to controls with the same or similar values of the propensity score (both of these approaches were used by Dehejia and Wahba, 1999). Alternately, a model-based or non-parametric estimate of $E[\mathrm{Y}_i|p(\mathrm{X}_i),\mathrm{D}_i]$ can be substituted for these conditional mean functions and the outer expectation replaced with a sum (as in Heckman, Ichimura, and Todd, 1998).

The somewhat niftier weighting approach to propensity-score estimation skips the cumbersome matching step by exploiting the fact that the CIA implies $E\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)}\right] = E[\mathrm{Y}_{1i}]$ and $E[\frac{\mathrm{Y}_i(1-\mathrm{D}_i)}{(1-p(\mathrm{X}_i))}] = E[\mathrm{Y}_{0i}]$. Therefore, given a scheme for estimating $p(\mathrm{X}_i)$, we can construct estimates of the average treatment effect from the sample analog of

$$\begin{aligned} E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}] &= E\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)} - \frac{\mathrm{Y}_i(1-\mathrm{D}_i)}{1-p(\mathrm{X}_i)}\right] \\ &= E\left[\frac{(\mathrm{D}_i - p(\mathrm{X}_i))\mathrm{Y}_i}{p(\mathrm{X}_i)(1-p(\mathrm{X}_i))}\right]. \end{aligned} \tag{3.3.11}$$

This last expression is an estimand of the form suggested by Newey (1990) and Robins, Mark, and Newey

(1992). We can similarly calculate the effect of treatment on the treated from the sample analog of:

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1] = E\left[\frac{(\mathrm{D}_i - p(\mathrm{X}_i))\mathrm{Y}_i}{(1 - p(\mathrm{X}_i))P(\mathrm{D}_i)}\right]. \tag{3.3.12}$$

The idea that you can correct for non-random sampling by weighting by the reciprocal of the probability of selection dates back to Horvitz and Thompson (1952). Of course, to make this approach feasible, and for the resulting estimates to be consistent, we need a consistent estimator for $p(\mathrm{X}_i)$

The Horvitz-Thompson version of the propensity-score approach is appealing since the estimator is essentially automated, with no cumbersome matching required. The Horvitz-Thompson approach also highlights the close link between propensity-score matching and regression, much as discussed for covariate matching in section 3.3.1. Consider again the regression estimand, δ_R, for the population regression of Y_i on D_i, controlling for a saturated model for covariates. This estimand can be written

$$\delta_R = \frac{E[(\mathrm{D}_i - p(\mathrm{X}_i))\mathrm{Y}_i]}{E[p(\mathrm{X}_i)(1 - p(\mathrm{X}_i))]}. \tag{3.3.13}$$

The two Horvitz-Thompson matching estimands and the regression estimand are all members of the class of weighted average estimands considered by Hirano, Imbens, and Ridder (2003):

$$E\left\{g(\mathrm{X}_i)\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)} - \frac{\mathrm{Y}_i(1 - \mathrm{D}_i)}{(1 - p(\mathrm{X}_i))}\right]\right\}, \tag{3.3.14}$$

where $g(\mathrm{X}_i)$ is a known weighting function (To go from estimand to estimator, replace $p(\mathrm{X}_i)$ with a consistent estimator, and expectations with sums). For the average treatment effect, set $g(\mathrm{X}_i) = 1$; for the effect on the treated, set $g(\mathrm{X}_i) = \frac{p(\mathrm{X}_i)}{P(\mathrm{D}_i)}$; and for regression set

$$g(\mathrm{X}_i) = \frac{p(\mathrm{X}_i)(1 - p(\mathrm{X}_i))}{E[p(\mathrm{X}_i)(1 - p(\mathrm{X}_i))]}.$$

This similarity highlights once again the fact that regression and matching—including propensity score matching—are not really different animals, at least not until we specify a model for the propensity score.

A big question here is how best to model and estimate $p(\mathrm{X}_i)$, or how much smoothing or stratification to use when estimating $E[\mathrm{Y}_i|p(\mathrm{X}_i),\mathrm{D}_i]$, especially if the covariates are continuous The regression analog of this question is how to parametrize the control variables (e.g., polynomials or main effects and interaction terms if the covariates are coded as discrete). The answer to this is inherently application-specific. A growing empirical literature suggests that a Logit model for the propensity score with a few polynomial terms in continuous covariates works well in practice, though this cannot be a theorem (see, e.g., Dehejia and Wahba, 1999).

A developing theoretical literature has produced some thought-provoking theorems on efficient use of the

propensity score. First, from the point of view of asymptotic efficiency, there is usually a cost to matching on the propensity score instead of full covariate matching. We can get lower asymptotic standard errors by matching on any covariate that explains outcomes, whether or not it turns up in the propensity score. This we know from Hahn's (1998) investigation of the maximal precision that it is possible to obtain for estimates of treatment effects under the CIA, with and without knowledge of the propensity score. For example, in Angrist (1998), there is an efficiency gain from matching on year of birth, even if the probability of serving in the military is unrelated to birth year, because earnings are related to birth year. A regression analog for this point is the result that even in a scenario with no omitted variables bias, the long regression generates more precise estimates of the coefficients on the variables included in a short regression whenever these variables have some predictive power for outcomes because these covariates lead to a smaller residual variance (see Section 3.1.3).

Hahn's (1998) results raise the question of why we should ever bother with estimators that use the propensity score. A philosophical argument is that the propensity score rightly focuses researcher attention on models for treatment assignment, something about which we may have reasonably good information, instead of the typically more complex and mysterious process determining outcomes. This view seems especially compelling when treatment assignment is the outcome of human institutions or government regulations while the process determining outcomes is more anonymous (e.g., a market). For example, in a time series evaluation of the causal effects of monetary policy, Angrist and Kuersteiner (2004) argue that we know more about how the Federal Reserve sets interests rates than about the process determining GDP. In the same spirit, it may also be easier to validate a model for treatment assignment than to validate a model for outcomes (see, e.g., Rosenbaum and Rubin, 1985, for a version of this argument).

A more precise though purely statistical argument for using the propensity score is laid out in Angrist and Hahn (2004). This paper shows that even though there is no asymptotic efficiency gain from the use of estimators based on the propensity score, there will often be a gain in precision in finite samples. Since all real data sets are finite, this result is empirically relevant. Intuitively, if the covariates omitted from the propensity score explain little of the variation in outcomes (in a purely statistical sense), it may then be better to ignore them than to bear the statistical burden imposed by the need to estimate their effects. This is easy to see in studies using data sets such as the NLSY where there are hundreds of covariates that might predict outcomes. In practice, we focus on a small subset of all possible covariates. This subset is chosen with an eye to what predicts treatment as well as outcomes.

Finally, Hirano, Imbens, and Ridder (2003) provide an alternative asymptotic resolution of the "propensity score paradox" generated by Hahn's (1998) theorems. They show that even though estimates of treatment effects based on a known propensity score are inefficient, for models with continuous covariates, a Horvitz-Thompson-type weighting estimator is efficient when weighting uses a *non-parametric estimate* of the score. The fact that the propensity score is estimated and the fact that it is estimated non-parametrically

are both key for the Hirano, Imbens, and Ridder conclusions.

Do the Hirano, Imbens, and Ridder (2003) results resolve the propensity-score paradox? For the moment, we prefer the finite-sample resolution given by Angrist and Hahn (2004). Their results highlight the fact that it is the researchers' willingness to impose some restrictions on the score which gives propensity-score-based inference its conceptual and statistical power. In Angrist (1998), for example, an application with high-dimensional though discrete covariates, the unrestricted non-parametric estimator of the score is just the empirical probability of treatment in each covariate cell. With this nonparametric estimator plugged in for $p(\mathrm{X}_i)$, it's straightforward to show that the sample analogs of (3.3.11) and (3.3.12) are algebraically equivalent to the corresponding full-covariate matching estimators. Hence, it's no surprise that score-based estimation comes out efficient, since full-covariate matching is the asymptotically efficient benchmark. An essential element of propensity score methods is the use of prior knowledge for dimension reduction. The statistical payoff is an improvement in finite-sample behavior. If you're not prepared to smooth, restrict, or otherwise reduce the dimensionality of the matching problem in a manner that has real empirical consequences, then you might as well go for full covariate matching or saturated regression control.

3.3.3 Propensity-Score Methods vs. Regression

Propensity-score methods shift attention from the estimation of $E[\mathrm{Y}_i|\mathrm{X}_i,\mathrm{D}_i]$ to the estimation of the propensity score, $p(\mathrm{X}_i) \equiv E[\mathrm{D}_i|\mathrm{X}_i]$. This is attractive in applications where the latter is easier to model or motivate. For example, Ashenfelter (1978) showed that participants in government-funded training programs often have suffered a marked pre-program dip in earnings, a pattern found in many later studies. If this dip is the only thing that makes trainees special, then we can estimate the causal effect of training on earnings by controlling for past earnings dynamics. In practice, however, it's hard to match on earnings dynamics since earnings histories are both continuous and multi-dimensional. Dehejia and Wahba (1999) argue in this context that the causal effects of training programs are better estimated by conditioning on the propensity score than by conditioning on the earnings histories themselves.

The propensity-score estimates reported by Dehejia and Wahba are remarkably close to the estimates from a randomized trial that constitute their benchmark. Nevertheless, we believe regression should be the starting point for most empirical projects. This is not a theorem; undoubtedly, there are circumstances where propensity score matching provides more reliable estimates of average causal effects. The first reason we don't find ourselves on the propensity-score bandwagon is practical: there are many details to be filled in when implementing propensity-score matching - such as how to model the score and how to do inference - these details are not yet standardized. Different researchers might therefore reach very different conclusions, even when using the same data and covariates. Moreover, as we've seen with the Horvitz-Thompson estimands, there isn't very much theoretical daylight between regression and propensity-score weighting. If the regression model for covariates is fairly flexible, say, close to saturated, regression can be seen as a type

of propensity-score weighting, so the difference is mostly in the implementation. In practice you may be far from saturation, but with the right covariates this shouldn't matter.

The face-off between regression and propensity-score matching is illustrated here using the same National Supported Work (NSW) sample featured in Dehejia and Wahba (1999).[25] The NSW is a mid-1970s program that provided work experience to a sample with weak labor-force attachment. Somewhat unusually for it's time, the NSW was evaluated in a randomized trial. Lalonde's (1986) path-breaking analysis compared the results from the NSW randomized study to econometric results using non-experimental control groups drawn from the PSID and the CPS. He came away pessimistic because plausible non-experimental methods generated a wide range of results, many of which were far from the experimental estimates. Moreover, Lalonde argued, an objective investigator, not knowing the results of the randomized trial, would be unlikely to pick the best econometric specifications and observational control groups.

In a striking second take on the Lalonde (1986) findings, Dehejia and Wahba (1999) found that they could come close to the NSW experimental results by matching the NSW treatment group to observational control groups selected using the propensity score. They demonstrated this using various comparison groups. Following Dehejia and Wahba (1999), we look again at two of the CPS comparison groups, first, a largely unselected sample (CPS-1) and then a narrower comparison group selected from the recently unemployed (CPS-3).

Table 3.3.2 (a replication of Table 1 in Dehejia and Wahba, 1999) reports descriptive statistics for the NSW treatment group, the randomly selected NSW control group, and our two observational control groups. The NSW treatment group and the randomly selected NSW control groups are younger, less educated, more likely to be nonwhite, and have much lower earnings than the general population represented by the CPS-1 sample. The CPS-3 sample matches the NSW treatment group more closely but still shows some differences, particularly in terms of race and pre-program earnings.

Table 3.3.3 reports estimates of the NSW treatment effect. The dependent variable is annual earnings in 1978, a year or two after treatment. Rows of the table show results with alternative sets of controls: none; all the demographic variables in Table 3.3.2; lagged (1975) earnings; demographics plus lagged earnings; demographics and two lags of earnings. All estimates are from regressions of 1978 earnings on a treatment dummy plus controls (the raw treatment-control difference appears in the first row).

Estimates using the experimental control group, reported in column 1, are in the order of $1,600-1,800. Not surprisingly, these estimates vary little across specifications. In contrast, the raw earnings gap between NSW participants and the CPS-1 sample, reported in column 2, is roughly $-8,500, suggesting this comparison is heavily contaminated by selection bias. The addition of demographic controls and lagged earnings narrows the gap considerably; the estimated treatment effect reaches (positive) $800 in the last row. The results

[25] An similar but more extended propensity-score face-off appears in the exchange beween Smith and Todd (2005) and Dehejia (2005).

are even better in column 3, which uses the narrower CPS-3 comparison group. The characteristics of this group are much closer to the those of NSW participants; consistent with this, the raw earnings difference is only $-635. The fully-controlled estimate, reported in the last row, is close to $1,400, not far from the experimental treatment effect.

A drawback of the process taking us from CPS-1 to CPS-3 is the *ad hoc* nature of the rules used to construct the smaller and more carefully-selected CPS-3 comparison group. The CPS-3 selection criteria can be motivated by the NSW program rules, which favor individuals with low earnings and weak labor-force attachment, but in practice, there are many ways to implement this. We'd therefore like a more systematic approach to pre-screening. In a recent paper, Crump, Hotz, Imbens and Mitnik (2006) suggest that the propensity score be used for systematic sample-selection as a *precursor* to regression estimation. This contrasts with our earlier discussion of the propensity score as the basis for an *estimator*.

We implemented the Crump, *et al.* (2006) suggestion by first estimating the propensity score on a pooled NSW-treatment and observational-comparison sample, and then picking only those observations with $0.1 < p(\mathrm{X}_i) < 0.9$. In other words, the estimation sample is limited to observations with a predicted probability of treatment equal to at least 10 percent, but no more than 90 percent. This ensures that regressions are estimated with a sample including only covariate cells with there are at least a few treated and control observations. Estimation using screened samples therefore requires no extrapolation to cells without "common support", i.e. to cells where there is no overlap in the covariate distribution between treatment and controls. Descriptive statistics for samples screened on the score (estimated using the full set of covariates listed in the table) appear in the last two columns of Table 3.3.2. The covariate means in screened CPS-1 and CPS-3 are much closer to the NSW means in column 1 than are the covariate means from unscreened samples.

We explored the common-support screener further using alternative sets of covariates, but with the same covariates used for both screening and the estimation of treatment effects at each iteration. The resulting estimates are displayed in the final two columns of Table 3.3.3. Controlling for demographic variables or lagged earnings alone, these results differ little from those in columns 2-3. With both demographic variables and a single lag of earnings as controls, however, the screened CPS-1 estimates are quite a bit closer to the experimental estimates than are the unscreened results. Screened CPS-1 estimates with two lags of earnings remain close to the experimental benchmark. On the other hand, the common-support screener improves the CPS-3 results only slightly with a single lag of earnings and seems to be a step backward with two.

This investigation boosts our (already strong) faith in regression. Regression control for covariates does a good job of eliminating selection bias in the CPS-1 sample in spite of a huge baseline gap. Restricting the sample using our knowledge of program admissions criteria yields even better regression estimates with CPS-3, about as good as Dehejia and Wahba's (1999) propensity score matching results with two lags of earnings. Systematic pre-screening to enforce common support seems like a useful adjunct to regression

estimation with CPS-1, a large and coarsely-selected initial sample. The estimates in screened CPS-1 are as good as unscreened CPS-3. We note, however, that the standard errors for estimates using propensity-score-screened samples have not been adjusted to reflect sampling variance in our estimates of the score. An advantage of pre-screening using prior information, as in the step from CPS-1 to CPS-3, is that no such adjustment is necessary.

3.4 Regression Details

3.4.1 Weighting Regression

Few things are as confusing to applied researchers as the role of sample weights. Even now, 20 years post-Ph.D., we read the section of the Stata manual on weighting with some dismay. Weights can be used in a number of ways, and how they are used may well matter for your results. Regrettably, however, the case for or against weighting is often less than clear-cut, as are the specifics of how the weights should be programmed. A detailed discussion of weighting pros and cons is beyond the scope of this book. See Pfefferman (1993) and Deaton (1997) for two perspectives. In this brief subsection, we provide a few guidelines and a rationale for our approach to weighting.

A simple rule of thumb for weighting regression is use weights when they make it more likely that the regression you are estimating is close to the population target you are trying to estimate. If, for example, the target (or *estimand*) is the population regression function, and the sample to be used for estimation is non-random with sampling weights, w_i, equal to the inverse probability of sampling observation i, then it makes sense to use weighted least squares, weighting by w_i (for this you can use Stata *pweights or a SAS WEIGHT* statement). Weighting by the inverse sampling probability generates estimates that are consistent for the population regression function even if the sample you have to work with is not a simple random sample.

A related weighting scenario is grouped data. Suppose that you would like to regress Y_i on X_i in a random sample, presumably because you want to learn about the population regression vector $\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$. Instead of a random sample, however, you have data grouped at the level of X_i. That is, you have estimates of $E[\mathrm{Y}_i|\mathrm{X}_i = x]$ for each x, estimated using data from a random sample. Let this average be denoted $\bar{y}_x$, and suppose you also know n_x, where n_x/N is the relative frequency of x in the underlying random sample. As we saw in Section 3.1.2, the regression of $\bar{y}_x$ on x, weighted by n_x is the same as the random-sample regression. Therefore, if your goal is to get back to the microdata regression, it makes sense to weight by group size. We note, however, that macroeconomists, accustomed to working with published averages and ignoring the underlying microdata, might disagree, or perhaps take the point in principle but remain disinclined to buck tradition in their discipline, which favors the unweighted analysis of aggregates.

If, on the other hand, the rationale for weighting has something to do with heteroskedasticity, as in many

Table 3.3.2: Covariate means in the NSW and observational control samples

Variable	NSW		Full Samples		P-score Screened Samples	
	Treated	Control	CPS-1	CPS-3	CPS-1	CPS-3
	(1)	(2)	(3)	(4)	(5)	(6)
Age	25.82	25.05	33.23	28.03	25.63	25.97
Years of schooling	10.35	10.09	12.03	10.24	10.49	10.42
Black	0.84	0.83	0.07	0.20	0.96	0.52
Hispanic	0.06	0.11	0.07	0.14	0.03	0.20
Dropout	0.71	0.83	0.30	0.60	0.60	0.63
Married	0.19	0.15	0.71	0.51	0.26	0.29
1974 earninigs	2,096	2,107	14,017	5,619	2,821	2,969
1975 earnings	1,532	1,267	13,651	2,466	1,950	1,859
Number of Obs.	185	260	15,992	429	352	157

Notes: Adapted from Dehejia and Wahba (1999), Table 1. The samples in the first four columns are as described in Dehejia and Wahba (1999). The samples in the last two columns are limited to observations with a propensity score between .1 and .9. Propensity score estimates use all the covariates listed in the table.

Table 3.3.3: Regression estimates of NSW training effects using alternate controls

Specification	Full Samples			P-Score Screened Samples	
	NSW	CPS-1	CPS-3	CPS-1	CPS-3
	(1)	(2)	(3)	(4)	(5)
Raw Difference	1,794 (633)	-8,498 (712)	-635 (657)		
Demographic controls	1,670 (639)	-3,437 (710)	771 (837)	-3,361 (811) [139/497]	890 (884) [154/154]
1975 Earnings	1,750 (632)	-78 (537)	-91 (641)	no obs. [0/0]	166 (644) [183/427]
Demographics, 1975 Earnings	1,636 (638)	623 (558)	1,010 (822)	1,201 (722) [149/357]	1,050 (861) [157/162]
Demographics, 1974 and 1975 Earnings	1,676 (639)	794 (548)	1,369 (809)	1,362 (708) [151/352]	649 (853) [147/157]

Notes: The table reports regression estimates of training effects using the Dehejia-Wahba (1999) data with alternative sets of controls. The demographic controls are age, years of schooling, and dummies for Black, Hispanic, high school dropout, and married.

Standard Errors are reported in parentheses, Observation counts are reported in brackets [treated/control]

textbook discussions of weighting, we are even less sympathetic to weighting than the macroeconomists. The argument for weighting under heteroskedasticity goes roughly like this: suppose you are interested in a linear CEF, $E[\mathrm{Y}_i|\mathrm{X}_i] = \mathrm{X}_i'\beta$. The error term, defined as $e_i \equiv \mathrm{Y}_i - \mathrm{X}_i'\beta$, may be heteroskedastic. That is, the conditional variance function, $E[e_i^2|\mathrm{X}_i]$ need not be constant. In this case, while the population regression function is still equal to $E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$, the sample analog is inefficient. A more precise estimator of the linear CEF is weighted least squares, i.e., minimize the sum of squared errors weighted by an estimate of $E[e_i^2|\mathrm{X}_i]^{-1}$.

As noted in Section 3.1.3, an inherently heteroskedastic scenario is the LPM, where Y_i is a dummy variable. Assuming the CEF is in fact linear, as it will be if the model is saturated, then $P[\mathrm{Y}_i = 1|\mathrm{X}_i] = \mathrm{X}_i'\beta$ and therefore $E\left[e_i^2|\mathrm{X}_i\right] = \mathrm{X}_i'\beta\left(1 - \mathrm{X}_i'\beta\right)$, which is obviously a function of X_i. This is an example of model-based heteroskedasticity where in principle, the conditional variance function is easily constructed from estimates of the underlying regression function. The efficient weighted least squares estimator—a special case of generalized least squares (GLS)—is to weight by $\left[\mathrm{X}_i'\beta(1 - \mathrm{X}_i'\beta)\right]^{-1}$. In practice, because the CEF has been assumed to be linear, these weights can be estimated in a first pass by OLS.

There are two reason why we prefer not to weight in this case (though we would use a heteroskedasticity-consistent covariance matrix). First, in practice, the estimate of $E[e_i^2|\mathrm{X}_i]$ may not be very good. If the conditional variance model is a poor approximation and/or the estimates of it are very noisy (in the LPM, this might mean the CEF is not really linear), weighted least squares estimates may have worse finite-sample properties than unweighted estimates. The inferences you draw based on asymptotic theory may therefore be misleading, and the hoped for efficiency gain may not materialize[26]. Second, if the CEF is not linear, the weighted least squares estimator is no more likely to estimate the CEF than is the unweighted estimator. Moreover, the unweighted estimator still estimates something easy to interpret: it estimates the MMSE linear approximation to the population CEF.

Of course, the GLS estimator also provides some sort of approximation, but the nature of this approximation depends on the weights. At a minimum, this makes it harder to compare your results to estimates by other researchers, and opens up additional avenues for specification searches when results depend on weighting. Finally, an old caution comes to mind: "if it ain't broke, don't fix it." The interpretation of the population regression vector is unaffected by heteroskedasticity, so why worry about it? Any efficiency gain from weighting is likely to be modest, and incorrect or poorly estimated weights can do more harm than good.

3.4.2 Limited Dependent Variables and Marginal Effects

Many empirical studies involve variables that take on only a limited number of values. An example is the Angrist and Evans (1998) investigation of the effect of childbearing on female labor supply, discussed in

[26] Altonji and Segal (1996) discuss this point in a generalized method-of-moments context.

Section 3.4.2 in this chapter and in the chapter on instrumental variables, below. This study is concerned with the causal effects of childbearing on parents' work and earnings. Because childbearing is likely to be correlated with potential earnings, the study reports instrumental variables estimates based on sibling-sex composition and multiple births, as well as OLS estimates. Almost every outcome in this study is either binary (like employment status) or non-negative (like hours worked, weeks worked, and earnings). Should the fact that a dependent variable is limited affect empirical practice? Many econometrics textbooks argue that, while OLS is fine for continuous dependent variables, when the outcome of interest is a limited dependent variable (LDV), linear regression models are inappropriate and nonlinear models such as Probit and Tobit are preferred. In contrast, our view of regression as inheriting its legitimacy from the CEF makes LDVness seem less central.

As always, a useful benchmark is a randomized experiment, where regression is simply a treatment-control difference. Consider regressions of various outcome variables on a randomly assigned regressor that indicates one of the treatment groups in the Rand Health Insurance Experiment (HIE; Manning, et al, 1987). In this ambitious experiment, probably the most expensive in American social science, the Rand Corporation set up a small health insurance company that charged no premium. Nearly 6,000 participants in the study were randomly assigned to health insurance plans with different features.

One of the most important features of any insurance plan is the portion of health care costs the insured individual is expected to pay. The HIE randomly assigned individuals to many different plans. One plan provided entirely free care, while the others included various combinations of co-payments, expenditure caps, and deductibles so that patients covered some of their health care costs out-of-pocket. The main purpose of the experiment was to learn whether the use of medical care is sensitive to cost and, if so, whether this affects health. The HIE results showed that those offered free or low-cost medical care used more of it, but they were not, for the most part, any healthier as a result. These findings helped pave the way for cost-sensitive health insurance plans and managed care.

Most of the outcomes in the HIE are LDVs. These include dummies indicating whether an experimental subject incurred any medical expenditures or was hospitalized in a given year and non-negative outcomes such as the number of face-to-face doctor visits and gross annual medical expenses (whether paid by patient or insurer). The expenditure variable is zero for about 20 percent of the sample. Results for two of the HIE treatment groups are reproduced in Table 3.4.1, derived from the estimates reported in Table 2 of Manning, *et al.* (1987). Table 3.4.1 shows average outcomes in the free care and individual deductible groups. The latter group faced a deductible of $150 per person or $450 per family per year for outpatient care, after which all costs were covered (There was no charge for inpatient care). The overall sample size in these two groups was a little over 3,000.

To simplify the LDV discussion, suppose that the comparison between free care and deductible plans is

Table 3.4.1: Average outcomes in two of the HIE treatment groups

Plan	Face-to-face visits	Outpatient Expenses (1984$)	Admissions	Prob. Any Medical (%)	Prob. Any Inpatient (%)	Total Expenses (1984$)
Free	4.55	340	.128	86.8	10.3	749
	(.168)	(10.9)	(.0070)	(.817)	(.45)	(39)
Individual	3.02	235	.115	72.3	9.6	608
Deductible	(.171)	(11.9)	(.0076)	(1.54)	(.55)	(46)
Deductible	-1.53	-105	-0.013	-14.5	-0.7	-141
minus free	(.240)	(16.1)	(.0103)	(1.74)	(.71)	(60)

Notes: Adapted from Manning (1987), Table 2. All standard errors (shown in parentheses) are corrected for intertemporal and intrafamily correlations. Amounts are in June 1984 dollars. Visits are face-to-face contacts with MD, DO, or other health providers; excludes visits only for radiology, anesthesiology or pathology services. Visits and expenses exclude dental care and outpatient psychotherapy.

the only comparison of interest and that treatment was determined by simple random assignment.[27] Let $\text{D}_i = 1$ denote assignment to the deductible group. By virtue of random assignment, the difference in means between those with $\text{D}_i = 1$ and $\text{D}_i = 0$ identifies the effect of treatment on the treated. As in our earlier discussion of experiments (Chapter 2):

$$\begin{aligned} & E\left[\text{Y}_i|\text{D}_i = 1\right] - E\left[\text{Y}_i|\text{D}_i = 0\right] \\ = \; & E\left[\text{Y}_{1i}|\text{D}_i = 1\right] - E\left[\text{Y}_{0i}|\text{D}_i = 1\right] \\ = \; & E\left[\text{Y}_{1i} - \text{Y}_{0i}\right] \end{aligned} \tag{3.4.1}$$

because D_i is independent of potential outcomes. Also, as before, $E\left[\text{Y}_i|\text{D}_i = 1\right] - E\left[\text{Y}_i|\text{D}_i = 0\right]$ is the slope coefficient in a regression of Y_i on D_i.

Equation (3.4.1) suggests that the estimation of causal effects in experiments presents no special challenges whether Y_i is binary, non-negative, or continuously distributed. The interpretation of the right-hand side changes for different sorts of dependent variables, but you do not need to *do* anything special to get the average causal effect. For example, one of the HIE outcomes is a dummy denoting any medical expenditure.

[27] The HIE was considerably more complicated than described here. There were 14 different treatments, including assignment to a prepaid HMO-like service. The experimental design did not use simple random assignment, but rather a more complicated assignment scheme meant to ensure covariate balance acrosss groups.

Since the outcome here is a Bernoulli trial, we have

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}] = E[\mathrm{Y}_{1i}] - E[\mathrm{Y}_{0i}] = P[\mathrm{Y}_{1i} = 1] - P[\mathrm{Y}_{0i} = 1]. \tag{3.4.2}$$

This relation might affect the language we use to describe the results but not the underlying calculation. In the HIE, for example, comparisons across experimental groups, as on the left hand side of (3.4.1), show that 87 percent of those assigned to the free-care group used at least some care in a given year, while only 72 percent of those assigned to the deductible plan used care. The relatively modest $150 deductible therefore had a marked effect on use of care. The difference between these two rates, $-.15(s.e. = .017)$ is an estimate of $E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}]$, where Y_i is a dummy indicating any medical expenditure. Because the outcome here is a dummy variable, the average causal effect is also a causal effect on usage rates or probabilities.

Recognizing that the outcome variable here is a probability, suppose instead that you use Probit to fit the CEF in this case. No harm in trying! The Probit model is usually motivated by the assumption that participation is determined by a latent variable, Y_i^*, that satisfies

$$\mathrm{Y}_i^* = \beta_0^* + \beta_1^* \mathrm{D}_i - \nu_i, \tag{3.4.3}$$

where ν_i is distributed $N(0, \sigma^2)$. Note that this variable cannot be actual medical expenditure since expenditure is non-negative and therefore non-Normal, while Normally distributed variables are continuously distributed on the Real line and can therefore be negative. Given the latent index model,

$$\mathrm{Y}_i = 1[\mathrm{Y}_i^* > 0],$$

the CEF can be written

$$E[\mathrm{Y}_i|\mathrm{D}_i] = \Phi[\frac{\beta_0^* + \beta_1^* \mathrm{D}_i}{\sigma}],$$

where $\Phi[\cdot]$ is the Normal CDF. Therefore

$$E[\mathrm{Y}_i|\mathrm{D}_i] = \Phi[\frac{\beta_0^*}{\sigma}] + \{\Phi[\frac{\beta_0^* + \beta_1^*}{\sigma}] - \Phi[\frac{\beta_0^*}{\sigma}]\}\mathrm{D}_i.$$

This is a linear function of the regressor, D_i, so the slope coefficient in the regression of Y_i on D_i is exactly the difference in Probit fitted values, $\Phi[\frac{\beta_0^* + \beta_1^*}{\sigma}] - \Phi[\frac{\beta_0^*}{\sigma}]$. Note, however, that the *Probit Coefficients,* $\frac{\beta_0^*}{\sigma}$ and $\frac{\beta_1^*}{\sigma}$ do not give us the size of effect of D_i on participation until we feed them back into the Normal CDF (though they do have the right sign).

One of the most important outcomes in the HIE is gross medical expenditure, in other words, health care costs. Did subjects who faced a deductible use less care, as measured by the cost? In the HIE, the average difference in expenditures between the deductible and free-care groups was -141 dollars ($s.e. = 60$), about

19% of the expenditure level in the free-care group. This calculation suggests that making patients pay a portion of costs reduces expenditures quite a bit, though the estimate is not very precise.

Because expenditure outcomes are non-negative random variables, and sometimes equal to zero, their expectation can be written

$$E[\mathrm{Y}_i|\mathrm{D}_i] = E[\mathrm{Y}_i|\mathrm{Y}_i > 0, \mathrm{D}_i]P[\mathrm{Y}_i > 0|\mathrm{D}_i].$$

The difference in expenditure outcomes across treatment groups is

$$\begin{aligned} & E\left[\mathrm{Y}_i|\mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i|\mathrm{D}_i = 0\right] \qquad (3.4.4) \\ = \; & E\left[\mathrm{Y}_i|\mathrm{Y}_i > 0, \mathrm{D}_i = 1\right] P\left[\mathrm{Y}_i > 0|\mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i|\mathrm{Y}_i > 0, \mathrm{D}_i = 0\right] P\left[\mathrm{Y}_i > 0|\mathrm{D}_i = 0\right] \\ = \; & \underbrace{\left\{P\left[\mathrm{Y}_i > 0|\mathrm{D}_i = 1\right] - P\left[\mathrm{Y}_i > 0|\mathrm{D}_i = 0\right]\right\}}_{\text{participation effect}} E\left[\mathrm{Y}_i|\mathrm{Y}_i > 0, \mathrm{D}_i = 1\right] \\ & + \underbrace{\left\{E\left[\mathrm{Y}_i|\mathrm{Y}_i > 0, \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i|\mathrm{Y}_i > 0, \mathrm{D}_i = 0\right]\right\}}_{\text{COP effect}} P\left[\mathrm{Y}_i > 0|\mathrm{D}_i = 0\right]. \end{aligned}$$

So the overall difference in average expenditure can be broken up into two parts: the difference in the probability that expenditures are positive (often called a participation effect), and the difference in means conditional on participation, a conditional-on-positive (COP) effect. Again, however, this has no special implications for the estimation of causal effects; equation (3.4.1) remains true: the regression of Y_i on D_i gives the population average treatment effect for expenditures.

Good COP, Bad COP: Conditional-on-positive effects

Because the effect on a non-negative random variable like expenditure has two parts, some applied researchers feel they should look at these parts separately. In fact, many use a "two-part model," where the first part is an evaluation of effect on participation and the second part looks at the COP effects (see, e.g., Duan, *et al.*, 1983 and 1984 for such models applied to the HIE). The first part of (3.4.4) raises no special issues, because, as noted above, the fact that Y_i is a dummy means only that average treatment effects are also differences in probabilities. The problem with the two-part model is that the COP effects do not have a causal interpretation, even in a randomized trial. This is exactly the same selection problem raised in Section 3.2.3, on bad control.

To analyze the COP effect further, write

$$\begin{aligned} & E\left[\mathrm{Y}_i|\mathrm{Y}_i > 0, \mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i|\mathrm{Y}_i > 0, \mathrm{D}_i = 0\right] \qquad (3.4.5) \\ = \; & E\left[\mathrm{Y}_{1i}|\mathrm{Y}_{1i} > 0\right] - E\left[\mathrm{Y}_{0i}|\mathrm{Y}_{0i} > 0\right] \\ = \; & \underbrace{E\left[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{Y}_{1i} > 0\right]}_{\text{causal effect}} + \underbrace{\left\{E\left[\mathrm{Y}_{0i}|\mathrm{Y}_{1i} > 0\right] - E\left[\mathrm{Y}_{0i}|\mathrm{Y}_{0i} > 0\right]\right\}}_{\text{selection bias}}. \end{aligned}$$

This decomposition shows that the COP effect is composed of two terms: a causal effect for the subpopulation

that uses medical care when it is free and the difference in Y_{0i} between those who use medical care when it is free and those who use medical care when they have to pay something. This second term is a form of selection bias, though it is more subtle than the selection bias in Chapter 2.

Here selection bias arises because the experiment changes the *composition* of the group with positive expenditures. The $Y_{0i} > 0$ population probably includes some low-cost users who would opt out of care if they had to pay a deductible. In other words, it is larger and probably has lower costs on average than the $Y_{1i} > 0$ group. The selection bias term is therefore positive, with the result that COP effects are closer to zero than the negative causal effect, $E[Y_{1i} - Y_{0i} | Y_{1i} > 0]$. This is a version of the bad control problem from Section 3.2.3: in a causal-effects setting, $Y_i > 0$ is an outcome variable and therefore unkosher for conditioning unless the treatment has no effect on the likelihood that Y_i is positive.

One resolution of the non-causality of COP effects relies on censored regression models like Tobit. These models postulate a latent expenditure outcome for nonparticipants (e.g., Hay and Olsen, 1984). A traditional Tobit formulation for the expenditure problem stipulates that the observed Y_i is generated by

$$Y_i = 1[Y_i^* > 0] Y_i^*$$

where Y_i^* is a Normally distributed latent expenditure variable that can take on negative values. Because Y_i^* is not an LDV, Tobit proponents feel comfortable linking this to D_i with a traditional linear model, say, equation (3.4.3). In this case, β_1^* is the causal effect of D_i on latent expenditure, Y_i^*. This equation is defined for everyone, whether Y_i is positive or not. There is no COP-style selection problem if we are happy to study effects on Y_i^*.

But we are not happy with effects on Y_i^*. The first problem is that "latent health care expenditure" is a puzzling construct.[28] Health care expenditure really is zero for some people; this is not a statistical artifact or due to some kind of censoring. So the notion of latent and potentially negative Y_i^* is hard to grasp. There is no data on Y_i^* and there never will be. A second problem is that the link between the parameter β_1^* in the latent model and causal effects on the observed outcome, Y_i, turns on distributional assumptions about the latent variable. To establish this link we evaluate the expectation of Y_i given D_i to find

$$E[Y_i | D_i] = \Phi\left[\frac{\beta_0^* + \beta_1^* D_i}{\sigma}\right][\beta_0^* + \beta_1^* D_i] + \sigma\phi\left[\frac{\beta_0^* + \beta_1^* D_i}{\sigma}\right] \tag{3.4.6}$$

where σ is the standard deviation of ν_i (see, e.g. McDonald and Moffitt, 1980). This expression involves the assumed Normality and homoskedasticity of ν_i and the assumption that Y_i can be represented as $1[Y_i^* > 0]Y_i^*$, as well as the latent coefficients.

[28] A generalization of Tobit is the sample selection model, where the latent variable determining participation is not the same as the latent expenditure variable. See, e.g., Maddala (1983). The same conceptual problems related to the interpretation of effects on latent variables arise in the sample selection model as with Tobit.

The Tobit CEF provides us with an expression for a treatment effect on observed expenditure. Specifically,

$$\begin{aligned} & E\left[\mathrm{Y}_i|\mathrm{D}_i=1\right]-E\left[\mathrm{Y}_i|\mathrm{D}_i=0\right] \\ = & \left\{\Phi\left[\frac{\beta_0^*+\beta_1^*}{\sigma}\right]\left[\beta_0^*+\beta_1^*\right]+\sigma\phi\left[\frac{\beta_0^*+\beta_1^*}{\sigma}\right]\right\}-\left\{\Phi\left[\frac{\beta_0}{\sigma}\right]\left[\beta_0^*\right]+\sigma\phi\left[\frac{\beta_0^*}{\sigma}\right]\right\} \end{aligned} \tag{3.4.7}$$

a rather daunting expression. But since the only conditioning variable is a dummy variable, D_i, none of this is necessary for the estimation of $E[\mathrm{Y}_i|\mathrm{D}_i=1]-E[\mathrm{Y}_i|\mathrm{D}_i=0]$. The slope coefficient from an OLS regression of Y_i on D_i recovers the CEF difference on the left hand side of (3.4.7) whether or not you adopt a Tobit model to explain the underlying structure.

COP effects are sometimes motivated by a researcher's sense that when the outcome distribution has a mass point - that is, it piles up on particular values like zero - or a heavily skewed distribution, or both, then an analysis of effects on averages misses something. Analyses of effects on averages indeed miss some things, like changes in the probability of specific values, or a shift in quantiles away from the median. But why not look at these distribution effects directly? A sensible alternative to COP effects looks directly at effects on distributions or quantiles. Distribution outcomes include the likelihood that annual medical expenditures exceed zero, 100 dollars, 200 dollars, and so on. This puts $1[\mathrm{Y}_i > c]$ for different choices of c on the left-hand side of the regression of interest. Econometrically, these outcomes are all in the category of equation (3.4.2). The idea of looking directly at distribution effects with linear probability models is illustrated by Angrist (2001), in an analysis of the effects of childbearing on hours worked. Alternately, if quantiles provide a focal point, we can use quantile regression to model them. Chapter 7 discusses this idea in detail.

Do Tobit-type latent-variable models ever make sense? Yes, if the data you are working with are truly censored. True censoring means the latent variable has an empirical counterpart that is the outcome of primary interest. A leading example from labor economics is CPS earnings data, which topcodes (censors) very high values of earnings to protect respondent confidentiality. Typically, we're interested in the causal effect of schooling on earnings as it appears on respondents' tax returns, not their CPS-topcoded earnings. Chamberlain (1994) shows that in some years, CPS topcoding reduces the measured returns to schooling considerably, and proposes an adjustment for censoring based on a Tobit-style adaptation of quantile regression. The use of quantile regression to model censored data is also discussed in Chapter 7.[29]

[29] We should note that our favorite regression example - a regression of log wages on schooling - may have a COP problem since the sample of log wages naturally omits those with zero earnings. This leads to COP-style selection bias if education affects the probability of working. In practice, therefore, we focus on samples of prime-age males where participation rates are high and reasonably stable across schooling groups (e.g., white men aged 40-49 in Figure 3.1.1).

Covariates lead to nonlinearity

True censoring as with the CPS topcode is rare, a fact that leaves limited scope for constructive applications of Tobit-type models in applied work. At this point, however, we have to hedge a bit. Part of the neatness in the discussion of experiments comes from the fact that $E[\text{Y}_i|\text{D}_i]$ is necessarily a linear function of D_i so that regression and the CEF are one and the same. In fact, this CEF is linear for any function of Y_i, including the distribution indicators, $1[\text{Y}_i > c]$. In practice, of course, the explanatory variable of interest isn't always a dummy, and there are usually additional covariates in the CEF, in which case, $E[\text{Y}_i|\text{X}_i,\text{D}_i]$ is almost certainly nonlinear for LDVs. Intuitively, as predicted means get close to the dependent variable boundaries, say because some covariate cells are close to the boundaries, the derivatives of the CEF for LDVs get smaller (think, for example, of the how the Normal CDF flattens at extreme values).

The upshot is that in LDV models with covariates, regression need not fit the CEF perfectly. It remains true, however, that the underlying CEF has a causal interpretation if the CIA holds. And if the CEF has a causal interpretation, it seems fair to say that regression has a causal interpretation as well, because it still provides the MMSE approximation to the CEF. Moreover, if the model for covariates is saturated, then regression also estimates a weighted average treatment effect similar to (3.3.1) and (3.3.3). Likewise, if the regressor of interest is multi-valued or continuous, we get a weighted average derivative, as described by the formulas in subsection 3.3.1.

And yet, we don't often have enough data for the saturated-covariate regression specification to be very attractive. Regression will therefore miss some features of the CEF. For one thing, it may generate fitted values outside the LDV boundaries. This fact bothers some researchers and has certainly generated a lot of bad press for the linear probability model. One attractive feature of nonlinear models like Probit and Tobit is that they produce CEFs that respect LDV boundaries. In particular, Probit fitted values are always between zero and one, while Tobit fitted values are positive (this is not obvious from equation 3.4.6). We might therefore prefer nonlinear models on simple curve-fitting grounds.

Point conceded. It's important to emphasize, however, that the output from nonlinear models must be converted into *marginal effects* to be useful. Marginal effects are the (average) changes in CEF implied by a nonlinear model. Without marginal effects, it's hard to talk about the impact on observed dependent variables. Continuing to assume the regressor of interest is D_i, population average marginal effects can be constructed either by differencing

$$E\{E[\text{Y}_i|\text{X}_i,\text{D}_i = 1] - E[\text{Y}_i|\text{X}_i,\text{D}_i = 0]\},$$

or by differentiation: $E\left\{\frac{\partial E[\text{Y}_i|\text{X}_i,\text{D}_i]}{\partial \text{D}_i}\right\}$. Most people use derivatives when dealing with continuous or multi-valued regressors as well.

How close do OLS regression estimates come to the marginal effects induced by a nonlinear model like

Probit or Tobit? We first derive the marginal effects, and then show an empirical example. The Probit CEF for a model with covariates is

$$E[\mathrm{Y}_i|\mathrm{X}_i,\mathrm{D}_i] = \Phi\left[\frac{\mathrm{X}_i'\beta_0^* + \beta_1^*\mathrm{D}_i}{\sigma}\right].$$

The average finite difference is therefore

$$E\left\{\Phi\left[\frac{\mathrm{X}_i'\beta_0^* + \beta_1^*}{\sigma}\right] - \Phi\left[\frac{\mathrm{X}_i'\beta_0^*}{\sigma}\right]\right\}. \tag{3.4.8}$$

In practice, this can also be approximated by the average derivative,

$$E\left\{\Phi\left[\frac{\mathrm{X}_i'\beta_0^* + \beta_1^*\mathrm{D}_i}{\sigma}\right]\right\} \cdot (\beta_1^*/\sigma)$$

(Stata computes marginal effects both ways but defaults to (3.4.8) for dummy regressors).

Similarly, generalizing equation (3.4.6) to a model with covariates, we have

$$E[\mathrm{Y}_i|\mathrm{X}_i,\mathrm{D}_i] = \Phi\left[\frac{\mathrm{X}_i'\beta_0^* + \beta_1^*\mathrm{D}_i}{\sigma}\right]\left[\mathrm{X}_i'\beta_0^* + \beta_1^*\mathrm{D}_i\right] + \sigma\phi\left[\frac{\mathrm{X}_i'\beta_0^* + \beta_1^*\mathrm{D}_i}{\sigma}\right]$$

for a non-negative LDV. Tobit marginal effects are almost always cast in terms of the average derivative, which can be shown to be the surprisingly simple expression

$$E\left\{\phi\left[\frac{\mathrm{X}_i'\beta_0^* + \beta_1^*\mathrm{D}_i}{\sigma}\right]\right\} \cdot \beta_1^*. \tag{3.4.9}$$

See, e.g., Wooldridge (2006). One immediate implication of (3.4.9) is that the Tobit coefficient, β_1^* is always too big relative to the effect of D_i on Y_i. Intuitively, this is because - given the linear model for latent Y_i^* - the latent outcome always changes when D_i switches on or off. But real Y_i need not change: for many people, it's zero either way.

Table 3.4.2 compares regression and nonlinear marginal effects for a regression of female employment and hours of work, both LDVs, on measures of fertility. The estimates were constructed using one of the 1980 Census samples used by Angrist and Evans (1998) This sample includes married women aged 21-35 with at least two children. The childbearing variables consist of either a dummy indicating additional childbearing beyond two, or the total number of births. The covariates include linear terms in mothers' age, age at first birth, race dummies (black and Hispanic), and mother's education (dummies for high school graduates, some college, and college graduates). The covariate model is not saturated, rather there are linear effects and no interactions, so the underlying CEF in this example is surely nonlinear.

Probit marginal effects for the effect of a dummy variable indicating more than two children are indistinguishable from OLS estimates of the same relation. This can be seen in columns 2, 3, and 4 of Table 3.4.2,

the first row of which compares the estimates from different methods for the full 1980 sample. The OLS estimate of the effect of a third child is -.162, while the corresponding Probit marginal effects are -.163 and -.162. These were estimated using (3.4.8) in the first case and

$$E\left\{\Phi\left[\frac{\mathrm{X}_i'\beta_0^* + \beta_1^*}{\sigma}\right] - \Phi\left[\frac{\mathrm{X}_i'\beta_0^*}{\sigma}\right] |\mathrm{D}_i = 1\right\}$$

in the second (hence, a marginal effect on the treated).

Tobit marginal effects for the relation between fertility and hours worked are also quite close to the corresponding OLS estimates, though not indistinguishable. This can be seen in columns 5 and 6. Compare, for example, the Tobit estimates of -6.56 and -5.87 with the OLS estimate of -5.92 in column 2. Although one Tobit estimate is 10 percent larger in absolute value, this seems unlikely to be of substantive importance. The remaining columns of the table compare OLS to marginal effects for an ordinal childbearing variable instead of a dummy. These calculations all use derivatives to compute marginal effects (labeled MFX). Here too, the OLS and nonlinear marginal effects estimates are similar for both Probit and Tobit.

It is sometimes said that Probit models can be expected to generated marginal effects close to OLS when the fitted values are close to .5 because the nonlinear CEF is roughly linear in the middle. We therefore replicated the comparison of OLS and marginal effects in a subsample with relatively high average employment rates, non-white women over 30 who attended college and whose first birth was before age 20. Although the average employment rate is 83 percent in this group, the OLS estimates and marginal effects are again similar.

Table 3.4.2: Comparison of alternative estimates of the effect of childbearing on LDVs

		Right-hand side variable								
		More than two children					Number of children			
	Mean	OLS	Probit		Tobit		OLS	Probit MFX	Tobit MFX	
			Avg effect, full sample	Avg effect on treated	Avg effect, full sample	Avg effect on treated		Avg effect, full sample	Avg effect, full sample	Avg effect on treated
Dependent variable										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Full Sample										
Employment	.528	-.162	-.163	-.162	-	-	-.113	-.114	-	-
	(.499)	(.002)	(.002)	(.002)			(.001)	(.001)		
Hours worked	16.7	-5.92	-	-	-6.56	-5.87	-4.07	-	-4.66	-4.23
	(18.3)	(.074)			(.081)	(.073)	(.047)		(.054)	(.049)
Panel B: Non-white College Attendees over 30, first birth before age 20										
Employment	.832	-.061	-.064	-.070	-	-	-.054	-.048	-	-
	(.374)	(.028)	(.028)	(.031)			(.016)	(.013)		
Hours worked	30.8	-4.69	-	-	-4.97	-4.90	-2.83	-	-3.20	-3.15
	(16.0)	(1.18)			(1.33)	(1.31)	(.645)		(.670)	(.659)

Notes: The table reports OLS estimates, average treatment effects, and marginal effects (MFX) for the effect of childbearing on mothers' labor supply. The sample in Panel A includes 254,654 observations and is the same as the married-women-1980-Census sample used by Angrist and Evans (1998). Covariates include age, age at first birth, and dummies for boys at first and second birth. The sample in Panel B includes 746 nonwhites with at least some college aged over 30 whose first birth was before age 20. Standard deviations are reported in parentheses in column 1. Standard errors are shown in parentheses in other columns. The sample used to estimate average effects on the treated includes women with more than two children.

The upshot of this discussion is that while a nonlinear model may fit the CEF for LDVs more closely than a linear model, when it comes to marginal effects this probably matters little. This optimistic conclusion is not a theorem, but as in the empirical example here, it seems to be fairly robustly true.

Why then, should we bother with nonlinear models and marginal effects? One answer is that the marginal effects are easy enough to compute now that they are automated in packages like Stata. But there are a number of decisions to make along the way (e.g., the weighting scheme, derivatives versus finite differences) while OLS is standardized. Nonlinear life also promises to get considerably more complicated when we start to think about IV and panel data. Finally, extra complexity comes into the inference step as well, since we need standard errors for marginal effects. The principle of Occam's razor advises, "Entities should not be multiplied unnecessarily." In this spirit, we quote our former teacher, Angus Deaton (1997), pondering the nonlinear regression function generated by Tobit-type models:

> Absent knowledge of F [the distribution of the errors], this regression function does not even identify the β's [Tobit coefficients] - see Powell (1989) - but more fundamentally, we should ask how it has come about that we have to deal with such an awkward, difficult, and non-robust object.

3.4.3 Why is Regression Called Regression and What Does Regression-to-the-mean Mean?

The term regression originates with Francis Galton's (1886) study of height. Galton, who worked with samples of roughly-normally-distributed data on parents and children, noted that the CEF of a child's height given his parents' height is linear, with parameters given by the bivariate regression slope and intercept. Since height is stationary (its distribution is not changing [much] over time), the bivariate regression slope is also the correlation coefficient, i.e., between zero and one.

The single regressor in Galton's set-up, x_i, is average parent height and the dependent variable, Y_i, is the height the of adult children. The regression slope coefficient, as always, is $\beta_1 = \frac{Cov(\mathrm{Y}_i, x_i)}{V(x_i)}$, and the intercept is $\alpha = E[\mathrm{Y}_i] - \beta_1 E[\mathrm{X}_i]$. But because height is not changing across generations, the mean and variance of Y_i and x_i are the same. Therefore,

$$\begin{aligned} \beta_1 &= \frac{Cov(\mathrm{Y}_i, x_i)}{V(x_i)} = \frac{Cov(\mathrm{Y}_i, x_i)}{\sqrt{V(x_i)}\sqrt{V(\mathrm{Y}_i)}} = \rho_{xy} \\ \alpha &= E[\mathrm{Y}_i] - \beta_1 E[\mathrm{X}_i] = \mu(1-\beta_1) = \mu(1-\rho_{xy}) \end{aligned}$$

where ρ_{xy} is the intergenerational correlation coefficient in height and $\mu = E[\mathrm{Y}_i] = E[\mathrm{X}_i]$ is population average height. From this we get the linear CEF

$$E[\mathrm{Y}_i|x_i] = \mu(1-\rho_{xy}) + \rho_{xy}x_i,$$

so the height of a child given his parents' height is therefore a weighted average of his parents' height and the population average height. The child of tall parents will therefore not be as tall as they are, on average. Likewise, for the short. To be specific, Pischke, who is 6' 3", can expect his children to be tall, though not as tall as he is. Thankfully, however, Angrist, who is 5'6", can expect his children to be taller than he is. Galton called this property, "regression toward mediocrity in hereditary stature." Today, we call this "regression to the mean."

Galton, who was Charles Darwin's cousin, is also remembered for having founded the Eugenics Society, dedicated to breeding better people. Indeed, his interest in regression came largely from this quest. We conclude from this that the value of scientific ideas should not be judged by their author's politics.

Galton does not seem to have shown much interest in multiple regression, our chief concern in this chapter. Indeed, the regressions in Galton's work are mechanical properties of distributions of stationary random variables, almost identities, and certainly not causal. Galton, would have said so himself because he objected to the Lamarckian idea (later promoted in Stalin's Russia) that acquired traits could be inherited.

The idea that regression can be used for statistical control satisfyingly originates in an inquiry into the determinants of poverty rates by George Udny Yule (1899). Yule, a statistician and student of Karl Pearson's (Pearson was Galton's protégé) realized that Galton's regression coefficient could be extended to multiple variables by solving the least squares normal equations that had been derived long before by Legendre and Gauss. Yule's (1899) paper appears to be the first publication containing multivariate regression estimates. His model links changes in poverty rates in an area to changes in the administration of the English Poor Laws, while controlling for population growth and the age distribution in the area. He was particularly interested in whether out-relief, the practice of providing income support for poor people without requiring them to move to the poorhouse, did not itself contribute to higher poverty rates. This is a well-defined causal question of a sort that still occupies us today.[30]

Finally, we note that the history of regression is beautifully detailed in the book by Steven Stigler (1986). Stigler is a famous statistician at the University of Chicago, but not quite as famous as his father, the economist and Nobel laureate, George Stigler.

3.5 Appendix: Derivation of the average derivative formula

Begin with the regression of Y_i on S_i :

$$\frac{Cov(\mathrm{Y}_i, \mathrm{S}_i)}{V(\mathrm{S}_i)} = \frac{E[h(\mathrm{S}_i)(\mathrm{S}_i - E[\mathrm{S}_i])]}{E[\mathrm{S}_i(\mathrm{S}_i - E[\mathrm{S}_i])]}.$$

[30] Yule's first applied paper on the poor laws was published in 1895 in the *Economic Journal*, where Pischke is proud to serve as co-editor. The theory of multiple regression that goes along with this appears in Yule (1897).

Let $\kappa_{-\infty} = \lim_{t \to -\infty} h(t)$. By the fundamental theorem of calculus, we have:

$$h(\mathrm{S}_i) = \kappa_{-\infty} + \int_{-\infty}^{\mathrm{S}_i} h'(t)\, dt.$$

Substituting for $h(\mathrm{S}_i)$, the numerator becomes

$$E[h(\mathrm{S}_i)(\mathrm{S}_i - E[\mathrm{S}_i])] = \int_{-\infty}^{+\infty} \int_{-\infty}^{s} h'(t)\,(s - E[\mathrm{S}_i]) g(s) dt ds$$

where $g(s)$ is the density of S_i at s. Reversing the order of integration, we have

$$E[h(\mathrm{S}_i)(\mathrm{S}_i - E[\mathrm{S}_i])] = \int_{-\infty}^{+\infty} h'(t) \int_{t}^{+\infty} (s - E[\mathrm{S}_i]) g(s) ds dt.$$

The inner integral is easily seen to be equal to $\mu_t \equiv \{E[\mathrm{S}_i|\mathrm{S}_i \geq t] - E[\mathrm{S}_i|\mathrm{S}_i < t]\}\{P(\mathrm{S}_i \geq t)[1 - P(\mathrm{S}_i \geq t)\}$, which is clearly non-negative. Setting $\mathrm{S}_i = \mathrm{Y}_i$, the denominator can similarly be shown to be the integral of these weights. We therefore have a weighted average derivative representation of the bivariate regression coefficient, $\frac{Cov(\mathrm{Y}_i, \mathrm{S}_i)}{V(\mathrm{S}_i)}$, equation (3.3.8) in the text. A similar formula for a regression with covariates, X_i, is derived in the appendix to Angrist and Krueger (1999).

Chapter 4

Instrumental Variables in Action: Sometimes You Get What You Need

> Anything that happens, happens.
> Anything that, in happening, causes something else to happen,
> causes something else to happen.
> Anything that, in happening,
> causes itself to happen again, happens again.
> It doesn't necessarily do it in chronological order, though.
>
> Douglas Adams, *Mostly Harmless* (1995)

Two things distinguish the discipline of Econometrics from our older sister field of Statistics. One is a lack of shyness about causality. Causal inference has always been the name of the game in applied econometrics. Statistician Paul Holland (1986) cautions that there can be "no causation without manipulation," a maxim that would seem to rule out causal inference from non-experimental data. Less thoughtful observers fall back on the truism that "correlation is not causality." Like most people who work with data for a living, we believe that correlation can sometimes provide pretty good evidence of a causal relation, even when the variable of interest has not been manipulated by a researcher or experimenter. [1]

The second thing that distinguishes us from most statisticians—and indeed most other social scientists—is an arsenal of statistical tools that grew out of early econometric research on the problem of how to estimate the parameters in a system of linear simultaneous equations. The most powerful weapon in this arsenal is the method of Instrumental Variables (IV), the subject of this chapter. As it turns out, IV does more than allow us to consistently estimate the parameters in a system of simultaneous equations, though it allows us

[1] Recent years have seen an increased willingness by statisticians to discuss statistical models for observational data in an explicitly causal framework; see, for example, Freedman's (2005) review.

to do that as well.

Studying agricultural markets in the 1920s, the father and son research team of Phillip and Sewall Wright were interested in a challenging problem of causal inference: how to estimate the slope of supply and demand curves when observed data on prices and quantities are determined by the intersection of these two curves. In other words, equilibrium prices and quantities—the only ones we get to observe—solve these two stochastic equations at the same time. Upon which curve, therefore, does the observed scatterplot of prices and quantities lie? The fact that population regression coefficients do not capture the slope of any one equation in a set of simultaneous equations had been understood by Phillip Wright for some time. The IV method, first laid out in Wright (1928), solves the statistical simultaneous equations problem by using variables that appear in one equation to shift this equation and trace out the other. The variables that do the shifting came to be known as *instrumental variables* (Reiersol, 1941).

In a separate line of inquiry, IV methods were pioneered to solve the problem of bias from measurement error in regression models[2]. One of the most important results in the statistical theory of linear models is that a regression coefficient is biased towards zero when the regressor of interest is measured with random errors (to see why, imagine the regressor contains only random error; then it will be uncorrelated with the dependent variable, and hence the regression of $\texttt{Y}_i$ on this variable will be zero). Instrumental variables methods can be used to eliminate this sort of bias.

Simultaneous equations models (SEMs) have been enormously important in the history of econometric thought. At the same time, few of today's most influential applied papers rely on an orthodox SEM framework, though the technical language used to discuss IV still comes from this framework. Today, we are more likely to find IV used to address measurement error problems than to estimate the parameters of an SEM. Undoubtedly, however, the most important contemporary use of IV is to solve the problem of omitted variables bias. IV solves the problem of missing or unknown control variables, much as a randomized trial obviates the need for extensive controls in a regression.[3]

4.1 IV and causality

We like to tell the IV story in two iterations, first in a restricted model with constant effects, then in a framework with unrestricted heterogeneous potential outcomes, in which case causal effects must also be heterogeneous. The introduction of heterogeneous effects enriches the interpretation of IV estimands, without changing the mechanics of the core statistical methods we are most likely to use in practice (typically, two-stage least squares). An initial focus on constant effects allows us to explain the mechanics of IV with a

[2] Key historical references here are Wald (1940) and Durbin (1954), both discussed below.

[3] See Angrist and Krueger (2001) for a brief exposition of the history and uses of IV; Stock and Trebbi (2003) for a detailed account of the birth of IV; and Morgan (1990) for an extended history of econometric ideas, including the simultaneous equations model.

minimum of fuss.

To motivate the constant-effects setup as a framework for the causal link between schooling and wages, suppose, as before, that potential outcomes can be written

$$\mathrm{Y}_{si} \equiv f_i(s),$$

and that

$$f_i(s) = \pi_0 + \pi_1 s + \eta_i, \tag{4.1.1}$$

as in the introduction to regression in Chapter 3. Also, as in the earlier discussion, imagine that there is a vector of control variables, A_i, called "ability", that gives a selection-on-observables story:

$$\eta_i = A_i'\gamma + v_i,$$

where γ is again a vector of population regression coefficients, so that v_i and A_i are uncorrelated by construction. For now, the variables A_i, are assumed to be the only reason why η_i and S_i are correlated, so that

$$E[\mathrm{S}_i v_i] = 0.$$

In other words if A_i were observed, we would be happy to include it in the regression of wages on schooling; thereby producing a long regression that can be written

$$\mathrm{Y}_i = \alpha + \rho \mathrm{S}_i + A_i'\gamma + v_i. \tag{4.1.2}$$

Equation (4.1.2) is a version of the linear causal model, (3.2.9). The error term in this equation is the random part of potential outcomes, v_i, left over after controlling for A_i. This error term is uncorrelated with schooling by assumption. If this assumption turns out to be correct, the population regression of Y_i on S_i and A_i produces the coefficients in (4.1.2).

The problem we initially want to tackle is how to estimate the long-regression coefficient, ρ, when A_i is unobserved. Instrumental variables methods can be used to accomplish this when the researcher has access to a variable (the instrument, which we'll call Z_i), that is correlated with the causal variable of interest, S_i, but uncorrelated with any other determinants of the dependent variable. Here, the phrase "uncorrelated with any other determinants of the dependent variables" is like saying $Cov(\eta_i, \mathrm{Z}_i) = 0$, or, equivalently, Z_i is uncorrelated with both A_i and v_i. This statement is called an *exclusion restriction* since Z_i can be said to be excluded from the causal model of interest. The exclusion restriction is a version of the conditional independence assumption of the previous chapter, except that now it is the instrument which is independent of potential outcomes, instead of schooling itself (the "conditional" in conditional independence enters into

the discussion when we consider IV models with covariates).

Given the exclusion restriction, it follows from equation (4.1.2) that

$$\rho = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)}{Cov(\mathrm{S}_i, \mathrm{Z}_i)} = \frac{Cov(\mathrm{Y}_i, \mathrm{Z}_i)/V(\mathrm{Z}_i)}{Cov(\mathrm{S}_i, \mathrm{Z}_i)/V(\mathrm{Z}_i)}. \tag{4.1.3}$$

The second equality in (4.1.3) is useful because it's usually easier to think in terms of regression coefficients than in terms of covariances. The coefficient of interest, ρ, is the ratio of the population regression of Y_i on Z_i (the reduced form) to the population regression of S_i on Z_i (the first stage). The IV *estimator* is the sample analog of expression (4.1.3). Note that the IV *estimand* is predicated on the notion that the first stage is not zero, but this is something you can check in the data. As a rule, if the first stage is only marginally significantly different from zero, the resulting IV estimates are unlikely to be informative, a point we return to later.

It's worth recapping the assumptions needed for the ratio of covariances in (4.1.3) to equal the casual effect, ρ. First, the instrument must have a clear effect on S_i. This is the first stage. Second, the only reason for the relationship between Y_i and Z_i is the first-stage. For the moment, we're calling this second assumption the exclusion restriction, though as we'll see in the discussion of models with heterogeneous effects, this assumption really has two parts: the first is the statement that the instrument is as good as randomly assigned (i.e., independent of potential outcomes, conditional on covariates), while the second is that the instrument has no effect on outcomes other than through the first-stage channel.

So where can you find an instrumental variable? Good instruments come from institutional knowledge and your ideas about the processes determining the variable of interest. For example, the economic model of education suggests that educational attainment is determined by comparing the costs and benefits of alternative choices. Thus, one possible source of instruments for schooling is differences in costs due, say, to loan policies or other subsidies that vary independently of ability or earnings potential. A second source of variation in schooling is institutional constraints. A set of institutional constraints relevant for schooling are compulsory schooling laws. Angrist and Krueger (1991) exploit the variation induced by compulsory schooling in a paper that typifies the use of "natural experiments" to try to eliminate omitted variables bias

The starting point for the Angrist and Krueger (1991) quarter-of-birth strategy is the observation that most states required students to enter school in the calendar year in which they turn 6. School start age is therefore a function of date of birth. Specifically, those born late in the year are young for their grade. In states with a December 31st birthday cutoff, children born in the fourth quarter enter school shortly before they turn 6, while those born in the first quarter enter school at around age $6\frac{1}{2}$. Furthermore, because compulsory schooling laws typically require students to remain in school only until their 16th birthday, these groups of students will be in different grades or through a given grade to different degree, when they reach the legal dropout age. In essence, the combination of school start age policies and compulsory schooling laws

creates a natural experiment in which children are compelled to attend school for different lengths of time depending on their birthdays.

Angrist and Krueger looked at the relationship between educational attainment and quarter of birth using US census data. Panel A of Figure 4.1.1 (adapted from Angrist and Krueger, 2001) displays the education-quarter-of-birth pattern for men in the 1980 Census who were born in the 1930s. The figure clearly shows that men born earlier in the calendar year tend to have lower average schooling levels. Panel A of Figure 4.1.1 is a graphical representation of the first-stage. The first-stage in a general IV framework is the regression of the causal variable of interest on covariates and the instrument(s). The plot summarizes this regression because average schooling by year and quarter of birth is what you get for fitted values from a regression of schooling on a full set of year-of-birth and quarter-of-birth dummies.

Panel B of Figure 4.1.1 displays average earnings by quarter of birth for the same sample used to construct panel A. This panel illustrates what econometricians call the "reduced form" relationship between the instruments and the dependent variable. The reduced form is the regression of the dependent variable on any covariates in the model and the instrument(s). Panel B shows that older cohorts tend to have higher earnings, because earnings rise with work experience. The figure also shows that men born in early quarters almost always earned less, on average, than those born later in the year, even after adjusting for year of birth, which plays the role of an exogenous covariate in the Angrist and Krueger (1991) setup. Importantly, this reduced-form relation parallels the quarter-of-birth pattern in schooling, suggesting the two patterns are closely related. Because an individual's date of birth is probably unrelated to his or her innate ability, motivation, or family connections, it seems credible to assert that the only reason for the up-and-down quarter-of-birth pattern in earnings is indeed the up-and-down quarter-of-birth pattern in schooling. This is the critical assumption that drives the quarter-of-birth IV story.[4]

A mathematical representation of the story told by Figure 4.1.1 comes from the first-stage and reduced-form regression equations, spelled out below:

$$\mathrm{S}_i = \mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i + \xi_{1i} \tag{4.1.4a}$$

$$\mathrm{Y}_i = \mathrm{X}_i'\pi_{20} + \pi_{21}\mathrm{Z}_i + \xi_{2i} \tag{4.1.4b}$$

The parameter π_{11} in equation (4.1.4a) captures the first-stage effect of Z_i on S_i, adjusting for covariates,

[4] Other explanations are possible, the most likely being some sort of family background effect associated with season of birth (see, e.g., Bound, Jaeger, and Baker, 1995). Weighing against the possibility of omitted family background effects is the fact that the quarter of birth pattern in average schooling is much more pronounced at the schooling levels most affected by compulsory attendance laws. Another possible concern is a pure age-at-entry effect which operates through channels other than highest grade completed (e.g., achievement). The causal effect of age-at-entry on learning is difficult, if not impossible, to separate from pure age effects, as noted in Chapter 1). A recent study by Elder and Lubotsky (2008) argues that the evolution of putative age-at-entry effects over time is more consistent with effects due to age differences *per se* than to a within-school learning advantage for older students.

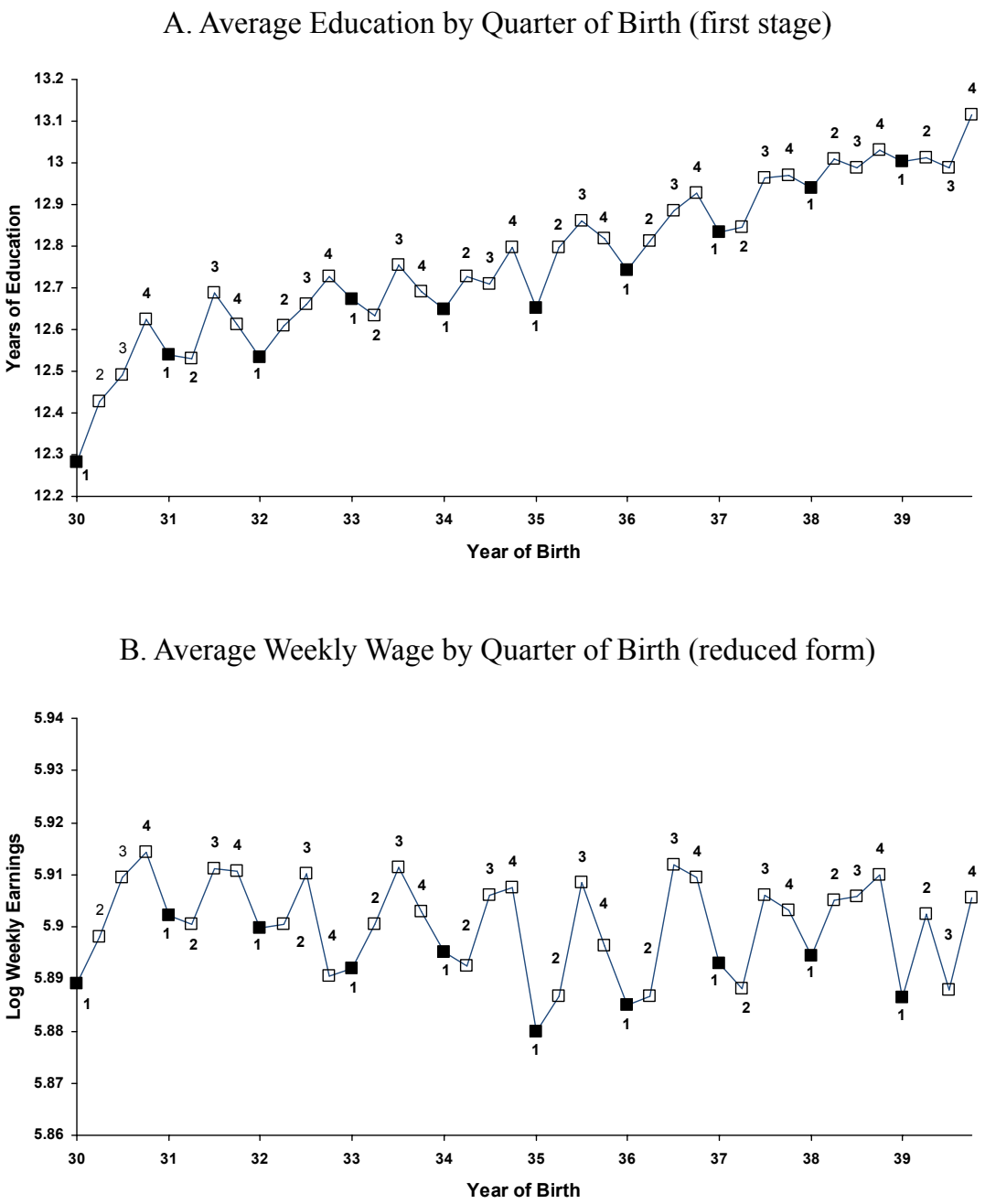

Figure 4.1.1: Graphical depiction of first stage and reduced form for IV estimates of the economic return to schooling using quarter of birth (from Angrist and Krueger 1991).

X_i. The parameter π_{21} in equation (4.1.4b) captures the reduced-form effect of Z_i on Y_i, adjusting for these same covariates. In the language of the SEM, the dependent variables in these two equations are said to be the *endogenous variables* (where they are determined jointly within the system) while the variables on the right-hand side are said to be the *exogenous variables* (determined outside the system). The instruments, Z_i, are a subset of the exogenous variables. The exogenous variables that are not instruments are said to be *exogenous covariates*. Although we're not estimating a traditional supply and demand system in this case, these SEM variable labels are still widely used in empirical practice.

The covariate-adjusted IV estimator is the sample analog of the ratio $\frac{\pi_{21}}{\pi_{11}}$. To see this, note that the denominators of the reduced-form and first-stage effects are the same. Hence, their ratio is

$$\rho = \frac{\pi_{21}}{\pi_{11}} = \frac{Cov(\mathrm{Y}_i, \tilde{z}_i)}{Cov(\mathrm{S}_i, \tilde{z}_i)}, \tag{4.1.5}$$

where $\tilde{z}_i$ is the residual from a regression of Z_i on the exogenous covariates, X_i. The right-hand side of (4.1.5) therefore swaps $\tilde{z}_i$ for Z_i in the general IV formula, (4.1.3). Econometricians call the sample analog of the left-hand side of equation (4.1.5) an Indirect Least Squares (ILS) estimator of ρ in the causal model with covariates,

$$\mathrm{Y}_i = \alpha' \mathrm{X}_i + \rho \mathrm{S}_i + \eta_i, \tag{4.1.6}$$

where η_i is the compound error term, $A_i'\gamma + v_i$[5]. It's easy to use equation (4.1.6) to confirm directly that $Cov(\mathrm{Y}_i, \tilde{z}_i) = \rho Cov(\mathrm{S}_i, \tilde{z}_i)$ since $\tilde{z}_i$ is uncorrelated with X_i by construction and with η_i by assumption. In Angrist and Krueger (1991), the instrument, Z_i, is quarter of birth (or dummies indicating quarters of birth) and the covariates are dummies for year of birth, state of birth, and race.

4.1.1 Two-Stage Least Squares

The reduced-form equation, (4.1.4b), can be derived by substituting the first stage equation, (4.1.4a), into the causal relation of interest, (4.1.6), which is also called a "structural equation" in simultaneous equations language. We then have:

$$\begin{aligned} \mathrm{Y}_i &= \alpha' \mathrm{X}_i + \rho[\mathrm{X}_i' \pi_{10} + \pi_{11} \mathrm{Z}_i + \xi_{1i}] + \eta_i \\ &= \mathrm{X}_i'[\alpha + \rho\pi_{10}] + \rho\pi_{11}\mathrm{Z}_i + [\rho\xi_{1i} + \eta_i] \\ &= \mathrm{X}_i'\pi_{20} + \pi_{21}\mathrm{Z}_i + \xi_{2i}, \end{aligned} \tag{4.1.7}$$

[5] For a direct proof that (4.1.5) equals ρ in (4.1.6), use (4.1.6) to substitute for Y_i in $\frac{Cov(\mathrm{Y}_i, \tilde{z}_i)}{Cov(\mathrm{S}_i, \tilde{z}_i)}$.

where $\pi_{20} \equiv \alpha + \rho\pi_{10}$, $\pi_{21} \equiv \rho\pi_{11}$, and $\xi_{2i} \equiv \rho\xi_{1i} + \eta_i$ in equation (4.1.4b). Equation (4.1.7) again shows why $\rho = \frac{\pi_{21}}{\pi_{11}}$. Note also that a slight re-arrangement of (4.1.7) gives

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho[\mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i] + \xi_{2i}, \tag{4.1.8}$$

where $[\mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i]$ is the population fitted value from the first-stage regression of S_i on X_i and Z_i. Because Z_i and X_i are uncorrelated with the reduced-form error, ξ_{2i}, the coefficient on $[\mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i]$ in the population regression of Y_i on X_i and $[\mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i]$ equals ρ.

In practice, of course, we almost always work with data from samples. Given a random sample, the first-stage fitted values in the population are consistently estimated by

$$\hat{s}_i = \mathrm{X}_i'\hat{\pi}_{10} + \hat{\pi}_{11}\mathrm{Z}_i,$$

where $\hat{\pi}_{10}$ and $\hat{\pi}_{11}$ are OLS estimates from equation (4.1.4a). The coefficient on $\hat{s}_i$ in the regression of Y_i on X_i and $\hat{s}_i$ is called the Two-Stage Least Squares (2SLS) estimator of ρ. In other words, 2SLS estimates can be constructed by OLS estimation of the "second-stage equation,"

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho\hat{s}_i + [\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)], \tag{4.1.9}$$

This is called 2SLS because it can be done in two steps, the first estimating $\hat{s}_i$ using equation (4.1.4a), and the second estimating equation (4.1.9). The resulting estimator is consistent for ρ because (a) first-stage estimates are consistent; and, (b) the covariates, X_i, and instruments, Z_i, are uncorrelated with both η_i and $(\mathrm{S}_i - \hat{s}_i)$.

The 2SLS name notwithstanding, we don't usually construct 2SLS estimates in two-steps. For one thing, the resulting standard errors are wrong, as we discuss later. Typically, we let specialized software routines (such as are available in SAS or Stata) do the calculation for us. This gets the standard errors right and helps to avoid other mistakes (see Section 4.6.1, below). Still, the fact that the 2SLS estimator can be computed by a sequence of OLS regressions is one way to remember why it works. Intuitively, conditional on covariates, 2SLS retains only the variation in S_i that is generated by quasi-experimental variation, i.e., generated by the instrument, Z_i.

2SLS is a many-splendored thing. For one, it is an instrumental variables estimator: the 2SLS estimate of ρ in (4.1.9) is the sample analog of $\frac{Cov(\mathrm{Y}_i, \hat{s}_i^*)}{Cov(\mathrm{S}_i, \hat{s}_i^*)}$, where $\hat{s}_i^*$ is the residual from a regression of $\hat{s}_i$ on X_i. This follows from the multivariate regression anatomy formula and the fact that $Cov(\mathrm{S}_i, \hat{s}_i^*) = V(\hat{s}_i^*)$. It is also easy to show that, in a model with a single endogenous variable and a single instrument, the 2SLS estimator is the same as the corresponding ILS estimator.[6]

[6]Note that $\hat{s}_i^* = \tilde{z}_i\hat{\pi}_{11}$, where $\tilde{z}_i$ is the residual from a regression of Z_i on X_i, so that the 2SLS estimator is therefore the

The link between 2SLS and IV warrants a bit more elaboration in the multi-instrument case. Assuming each instrument captures the same causal effect (a strong assumption that is relaxed below), we might want to combine these alternative IV estimates into a single more precise estimate. In models with multiple instruments, 2SLS provides just such a linear combination by combining multiple instruments into a single instrument. Suppose, for example, we have three instrumental variables, z_{1i}, z_{2i}, and z_{3i}. In the Angrist and Krueger (1991) application, these are dummies for first, second, and third-quarter births. The first-stage equation then becomes

$$\mathrm{s}_i = \mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{z}_{1i} + \pi_{12}\mathrm{z}_{2i} + \pi_{13}\mathrm{z}_{3i} + \xi_{1i}, \tag{4.1.10a}$$

while the 2SLS second stage is the same as (4.1.9), except that the fitted values are from (4.1.10a) instead of (4.1.4a). The IV interpretation of this 2SLS estimator is the same as before: the instrument is the residual from a regression of first-stage fitted values on covariates. The exclusion restriction in this case is the claim that all of the quarter of birth dummies in (4.1.10a) are uncorrelated with η_i in equation equation (4.1.6).

The results of 2SLS estimation of a schooling equation using three quarter-of-birth dummies, as well as other interactions, are shown in Table 4.1.1, which reports OLS and 2SLS estimates of models similar to those estimated by Angrist and Krueger (1991). Each column in the table contains OLS and 2SLS estimates of ρ from an equation like (4.1.6), estimated with different combinations of instruments and control variables. The OLS estimate in column 1 is from a regression of log wages with no control variables, while the OLS estimates in column 2 are from a model adding dummies for year of birth and state of birth as control variables. In both cases, the estimated return to schooling is around .075.

sample analog of $\frac{\left[\frac{Cov(\mathrm{y}_i,\tilde{z}_i)}{V(\tilde{z}_i)}\right]}{\hat{\pi}_{11}}$. But the sample analog of the numerator, $\frac{Cov(\mathrm{y}_i,\tilde{z}_i)}{V(\tilde{z}_i)}$, is the OLS estimate of π_{21} in the reduced form, (4.1.4b), while $\hat{\pi}_{11}$ is the OLS estimate of the first-stage effect, π_{11}, in (4.1.4a). Hence, 2SLS with a single instrument is ILS, i.e., the ratio of the reduced form-effect of the instrument to the corresponding first-stage effect where both the first-stage and reduced-form include covariates.

Table 4.1.1: 2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Years of education	0.075 (0.0004)	0.072 (0.0004)	0.103 (0.024)	0.112 (0.021)	0.106 (0.026)	0.108 (0.019)	0.089 (0.016)	0.061 (0.031)
Covariates:								
Age (in quarters)								✓
Age (in quarters) squared								✓
9 year of birth dummies		✓			✓	✓	✓	✓
50 state of birth dummies		✓			✓	✓	✓	✓
Instruments:			dummy for QOB=1	dummy for QOB=1 or QOB=2	dummy for QOB=1	full set of QOB dummies	full set of QOB dummies int. with year of birth dummies	full set of QOB dummies int. with year of birth dummies

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the the Angrist and Krueger (1991) 1980 Census sample. This sample includes native-born men, born 1930-1939, with positive earnings and non-allocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses.

The first pair of IV estimates, reported in columns 3 and 4, are from models without controls. The instrument used to construct the estimates in column 1 is a single dummy for first quarter births, while the instruments used to construct the estimates in column 2 are a pair of dummies indicating first and second quarter births. The standard error estimates range from .10 – .11. The results from models including year of birth and state of birth dummies as control variables are similar, not surprisingly, since quarter of birth is not closely related to either of these controls. Overall, the 2SLS estimates are mostly a bit larger than the corresponding OLS estimates. This suggests that the observed associated between schooling and earnings is not driven by omitted variables like ability and family background.

Column 7 in Table 4.1.1 shows the results of adding interaction terms to the instrument list. In particular, each specification adds interaction with 9 dummies for year of birth (the sample includes cohorts born 1930-39), for a total of 30 excluded instruments. The first stage equation becomes

$$\begin{aligned} \mathrm{S}_i &= \mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_{1i} + \pi_{12}\mathrm{Z}_{2i} + \pi_{13}\mathrm{Z}_{3i} \\ &\quad + \sum_j (\mathrm{B}_{ij}\mathrm{Z}_{1i})\kappa_{1j} + \sum_j (\mathrm{B}_{ij}\mathrm{Z}_{2i})\kappa_{2j} + \sum_j (\mathrm{B}_{ij}\mathrm{Z}_{3i})\kappa_{3j} + \xi_{1i} \end{aligned} \tag{4.1.10b}$$

where B_{ij} is a dummy equal to one if individual i was born in year j for j equal to 1931 – 39. The coefficients $\kappa_{1j}, \kappa_{2j}, \kappa_{3j}$ are the corresponding year-of-birth interactions. These interaction terms capture differences in the relation between quarter-of-birth and schooling across cohorts. The rationale for adding these interaction terms is an increase in precision that comes from increasing the first-stage R^2, which goes up because the quarter of birth pattern in schooling differs across cohorts. In this example, the addition of interaction terms to the instrument list leads to a modest gain in precision; the standard error declines from .0194 to .0161.[7]

The last 2SLS model reported in Table 4.1.1 includes controls for linear and quadratic terms in age-in-quarters in the list of covariates, X_i. In other words, someone who was born in the first quarter of 1930 is recorded as being 50 years old on census day (April 1), 1980, while someone born in the fourth quarter is recorded as being 49.25 years old. This finely coded age variable, entered into the model with a linear and quadratic term, provides a partial control for the fact that small differences age may be an omitted variable that confounds the quarter-of-birth identification strategy. As long as the effects of age are similarly smooth, the quadratic age-in-quarters model will pick them up.

This variation in the 2SLS set-up illustrates the inter-play between identification and estimation. For the 2SLS procedure to work, there must be some variation in the first-stage fitted values conditional on whatever control variables (covariates) are included in the model. If the first-stage fitted values are a linear combination of the included covariates, then the 2SLS estimate simply does not exist. In equation (4.1.9) this

[7] This gain may not be without cost, as the use of many additional instruments opens up the possibility of increased bias, an issue discussed in Chapter 8, below.

is manifest by perfect multicollinearity. 2SLS estimates with quadratic age exist. But the variability "left over" in the first-stage fitted values is reduced when the covariates include variables like age in quarters, that are closely related to the instruments (quarter of birth dummies). Because this variability is the primary determinant of 2SLS standard errors, the estimate in column 8 is markedly less precise than that in column 7, though it is still close to the corresponding OLS estimate.

Recap of IV and 2SLS Lingo

As we've seen, the *endogenous variables* are the dependent variable and the independent variable(s) to be instrumented; in a simultaneous equations model, endogenous variables are determined by solving a system of stochastic linear equations. To *treat an independent variable as endogenous* is to instrument it, i.e., to replace it with fitted values in the second stage of a 2SLS procedure. The independent endogenous variable in the Angrist and Krueger (1991) study is schooling. The *exogenous variables* include the *exogenous covariates* that are not instrumented and the instruments themselves. In a simultaneous equations model, exogenous variables are determined outside the system. The exogenous covariates in the Angrist and Krueger (1991) study are dummies for year of birth and state of birth. We think of exogenous covariates as controls. 2SLS aficionados live in a world of mutually exclusive labels: in any empirical study involving instrumental variables, the random variables to be studied are either dependent variables, independent endogenous variables, instrumental variables, or exogenous covariates. Sometimes we shorten this to: dependent and endogenous variables, instruments and covariates (fudging the fact that the dependent variable is also endogenous in a traditional SEM).

4.1.2 The Wald Estimator

The simplest IV estimator uses a single binary (0-1) instrument to estimate a model with one endogenous regressor and no covariates. Without covariates, the causal regression model is

$$\mathrm{Y}_i = \alpha + \rho \mathrm{S}_i + \eta_i, \tag{4.1.11}$$

where η_i and S_i may be correlated. Given the further simplification that Z_i is a dummy variable that equals 1 with probability p, we can easily show that

$$Cov(\mathrm{Y}_i, \mathrm{Z}_i) = \{E[\mathrm{Y}_i|\mathrm{Z}_i = 1] - E[\mathrm{Y}_i|\mathrm{Z}_i = 0]\}p(1-p),$$

with an analogous formula for $Cov(\mathrm{S}_i,\mathrm{Z}_i)$. It therefore follows that

$$\rho = \frac{E[\mathrm{Y}_i|\mathrm{Z}_i = 1] - E[\mathrm{Y}_i|\mathrm{Z}_i = 0]}{E[\mathrm{S}_i|\mathrm{Z}_i = 1] - E[\mathrm{S}_i|\mathrm{Z}_i = 0]}. \tag{4.1.12}$$

A direct route to this result uses (4.1.11) and the fact that $E[\eta_i|\mathrm{Z}_i] = 0$, so we have

$$E[\mathrm{Y}_i|\mathrm{Z}_i] = \alpha + \rho E[\mathrm{S}_i|\mathrm{Z}_i]. \tag{4.1.13}$$

Solving this equation for ρ produces (4.1.12).

Equation (4.1.12) is the population analog of the landmark Wald (1940) estimator for a bivariate regression with mismeasured regressors.[8] The *Wald estimator* is the sample analog of this expression. In our context, the Wald formula provides an appealingly transparent implementation of the IV strategy for the elimination of omitted variables bias. The principal claim that motivates IV estimation of causal effects is that the *only* reason for any relation between the dependent variable and the instrument is the effect of the instrument on the causal variable of interest. In the context of a binary instrument, it therefore seems natural to divide—or rescale—the reduced-form difference in means by the corresponding first-stage difference in means.

The Angrist and Krueger (1991) study using quarter of birth to estimate the economic returns to schooling shows the Wald estimator in action. Table 4.1.2 displays the ingredients behind a Wald estimate constructed using the 1980 census. The difference in earnings between men born in the first and second halves of the year is -.01349 (s.e.=.00337), while the corresponding difference in schooling is -.1514. The ratio of these two differences is a Wald estimate of the economic value of schooling in per-year terms. This comes out to be .0891 (s.e.=.021). Not surprisingly, this estimate is not too different from the 2SLS estimates in Table 4.1.1. The reason we should expect the Wald and 2SLS estimates to be similar is that they are both constructed from the same information: differences in earnings by season of birth.

The Angrist (1990) study of the effects of Vietnam-era military service on the earnings of veterans also shows the Wald estimator in action. In the 1960s and early 1970s, young men were at risk of being drafted for military service. Concerns about the fairness of US conscription policy led to the institution of a draft lottery in 1970 that was used to determine priority for conscription. A promising instrumental variable for Vietnam veteran status is therefore draft-eligibility, since this was determined by a lottery over birthdays. Specifically, in each year from 1970 to 1972, random sequence numbers (RSNs) were randomly assigned to each birth date in cohorts of 19-year-olds. Men with lottery numbers below an eligibility ceiling were eligible for the draft, while men with numbers above the ceiling could not be drafted. In practice, many draft-eligible men were still exempted from service for health or other reasons, while many men who were draft-exempt nevertheless volunteered for service. So veteran status was not completely determined by randomized draft-eligibility,

[8] As noted in the introduction to this chapter, measurement error in regressors tends to shrink regression coefficients towards zero. To eliminate this bias, Wald (1940) suggested that the data be divided in a manner independent of the measurement error, and the coefficient of interest estimated as a ratio of differences in means as in (4.1.12). Durbin (1954) showed that Wald's method of fitting straight lines is an IV estimator where the instrument is a dummy marking Wald's division of the data. Hausman (2001) provides an overview of econometric strategies for dealing with measurement error.

Table 4.1.2: Wald estimates of the returns to schooling using quarter of birth instruments

	(1) Born in the 1st or 2nd quarter of year	(2) Born in the 3rd or 4th quarter of year	(3) Difference (std. error) (1)-(2)
ln (weekly wage)	5.8916	5.9051	-0.01349 (0.00337)
Years of education	12.6881	12.8394	-0.1514 (0.0162)
Wald estimate of return to education			0.0891 (0.0210)
OLS estimate of return to education			0.0703 (0.0005)

Notes: Adapted from a re-analysis of Angrist and Krueger (1991) by Angrist and Imbens (1995). The sample includes native-born men with positive earnings from the 1930-39 birth cohorts in the 1980 Census 5 percent file. The sample size is 329,509.

but draft-eligibility provides a binary instrument highly correlated with Vietnam-era veteran status.

For white men who were at risk of being drafted in the 1970 draft lottery, draft-eligibility is clearly associated with lower earnings in years after the lottery. This is documented in Table 4.1.3, which reports the effect of randomized draft-eligibility status on average Social Security-taxable earnings in column 2. column 1 shows average annual earnings for purposes of comparison. For men born in 1950, there are significant negative effects of eligibility status on earnings in 1971, when these men were mostly just beginning their military service, and, perhaps more surprisingly, in 1981, ten years later. In contrast, there is no evidence of an association between draft-eligibility status and earnings in 1969, the year the lottery drawing for men born in 1950 was held but before anyone born in 1950 was actually drafted.

Because eligibility status was randomly assigned, the claim that the estimates in column 2 represent the effect of *draft-eligibility* on earnings seems uncontroversial. The information required to go from draft-eligibility effects to veteran-status effects is the denominator of the Wald estimator, which is the effect of draft-eligibility on the probability of serving in the military. This information is reported in column 3 of Table 4.1.3, which shows that draft-eligible men were almost 16 percentage points more likely to have served in the Vietnam era. The Wald estimate of the effect of military service on 1981 earnings, reported in column 4, amounts to about 15 percent of the mean. Effects were even larger in 1971 (in percentage terms), when affected soldiers were still in the army.

An important feature of the Wald/IV estimator is that the identifying assumptions are easy to assess and

Table 4.1.3: Wald estimates of the effects of military service on the earnings of white men born in 1950

	Earnings		Veteran Status		Wald Estimate of Veteran Effect
Earnings year	Mean	Eligibility Effect	Mean	Eligibility Effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	-435.8	0.267	0.159	-2,741
		(210.5)		(0.040)	(1,324)
1971	3,338	-325.9			-2050
		(46.6)			(293)
1969	2,299	-2.0			
		(34.5)			

Notes: Adapted from Angrist (1990), Tables 2 and 3. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Veteran status data are from the Survey of Program Participation. There are about 13,500 individuals in the sample.

interpret. Suppose D_i denotes Vietnam-era veteran status and Z_i indicates draft-eligibility. The fundamental claim justifying our interpretation of the Wald estimator as capturing the causal effect of D_i is that the only reason why $E[Y_i|Z_i]$ changes as Z_i changes is the variation in $E[D_i|Z_i]$. A simple check on this is to look for an association between Z_i and personal characteristics that should not be affected by D_i, for example, age, race, sex, or any other characteristic that was determined before D_i was determined. Another useful check is to look for an association between the instrument and outcomes in samples where there is no relationship between D_i and Z_i. If the only reason for draft-eligibility affects on earnings is veteran status, then draft-eligibility effects on earnings should be zero in samples where draft-eligibility status is unrelated to veteran status.

This idea is illustrated in the Angrist (1990) study of the draft lottery by looking at 1969 earnings, an estimate repeated in the last row of Table 4.1.3. It's comforting that the draft-eligibility treatment effect on 1969 earnings is zero since 1969 earnings predate the 1970 draft lottery. A second variation on this idea looks at the cohort of men born in 1953. Although there was a lottery drawing which assigned RSNs to the 1953 birth cohort in February of 1972, no one born in 1953 was actually drafted (the draft officially ended in July of 1973). The first-stage relationship between draft-eligibility and veteran status for men born in 1953 (defined using the 1952 lottery cutoff of 95) therefore shows only a small difference in the probability of serving by eligibility status. Importantly, there is also no significant relationship between earnings and draft-eligibility status for men born in 1953, a result that supports the claim that the only reason for draft-eligibility effects is military service.

We conclude the discussion of Wald estimators with a set of IV estimates of the effect of family size on mothers' employment and work. Like the schooling and military service studies, these estimates are used for illustration elsewhere in the book. The relationship between fertility and labor supply has long been of interest to labor economists, while the case for omitted variables bias in this context is clear: mothers with weak labor force attachment or low earnings potential may be more likely to have children than mothers with strong labor force attachment or high earnings potential. This makes the observed association between family size and employment hard to interpret since mothers who have big families may have worked less anyway. Angrist and Evans (1998) solve this omitted-variables problem using two instrumental variables, both of which lend themselves to Wald-type estimation strategies.

The first Wald estimator uses multiple births, an identification strategy for the effects of family size pioneered by Rosenzweig and Wolpin (1980). The twins instrument in Angrist and Evans (1998) is a dummy for a multiple third birth in a sample of mothers with at least two children. The twins first-stage is .625, an estimate reported in column 3 of Table 4.1.4. This means that 37.5 percent of mothers with two or more children would have had a third birth anyway; a multiple third birth increases this proportion to 1. The twins instrument rests on the idea that the occurrence of a multiple birth is essentially random, unrelated to potential outcomes or demographic characteristics.

The second Wald estimator in Table 4.1.4 uses sibling sex composition, an instrument motivated by the fact that American parents with two children are much more likely to have a third child if the first two are same-sex than if the sex-composition is mixed. This is illustrated in column 5 of Table 4.1.4, which shows that parents of same-sex sibling birth are 6.7 percentage points more likely to have a third birth (the probability of a third birth among parents with a mixed-sex sibship is .38). The same-sex instrument is based on the claim that sibling sex composition is essentially random and affects family labor supply solely by increasing fertility.

Twins and sex-composition instruments both suggest that the birth of a third child has a large effect on employment rates and on weeks and hours worked. Wald estimates using twins instruments show a precisely-estimate employment reduction of about .08, while weeks worked fall by 3.8 and hours per week fall by 3.4. These results, which appear in column 4 of Table 4.1.4, are smaller in absolute value than the corresponding OLS estimates reported in column 2. This suggests the latter are exaggerated by selection bias. Interestingly, the Wald estimates constructed using a same-sex dummy, reported in column 6, are larger than the twins estimates. The juxtaposition of twins and sex-composition instruments in Table 4.1.4 suggests that different instruments need not generate similar estimates of causal effects even if both are valid. We expand on this important point in Section 4.4. For now, however, we stick with a constant-effects framework.

Table 4.1.4: Wald estimates of labor supply effects

			IV Estimates using:			
			Twins		Sex-composition	
Dependent variable	Mean (1)	OLS (2)	First stage (3)	Wald estimates (4)	First stage (5)	Wald estimates (6)
Employment	0.528	-0.167	0.625	-0.083	0.067	-0.135
		(0.002)	(0.011)	(0.017)	(0.002)	(0.029)
Weeks worked	19.0	-8.05	"	-3.83	"	-6.23
		(0.09)	"	(0.758)	"	(1.29)
Hours/week	16.7	-6.02	"	-3.39	"	-5.54
		(0.08)	"	(0.637)	"	(1.08)

Note: The table reports OLS and Wald estimates of the effects of a third birth on labor supply using twins and sex-composition instruments. Data are from the Angrist and Evans (1998) extract including married women aged 21-35 with at least two children in the 1980 Census. OLS models include controls for mother's age, age at first birth, dummies for the sex of first and second births, and dummies for race.

4.1.3 Grouped Data and 2SLS

The Wald estimator is the mother of all instrumental variables estimators because more complicated 2SLS estimators can typically be constructed from an underlying set of Wald estimators. The link between Wald and 2SLS is grouped-data: 2SLS using dummy instruments is the same thing as GLS on a set of group means. GLS in turn can be understood as a linear combination of all the Wald estimators that can be constructed from pairs of means. The generality of this link might appear to be limited by the presumption that the instruments at hand are dummies. Not all instrumental variables are dummies, or even discrete, but this is not really important. For one thing, many credible instruments can be thought of as defining categories, such as quarter of birth. Moreover, instrumental variables that appear more continuous (such as draft lottery numbers, which range from 1-365) can usually be grouped without much loss of information (for example, a single dummy for draft-eligibility status, or dummies for groups of 25 lottery numbers).[9]

To explain the Wald/grouping/2SLS nexus more fully, we stick with the draft-lottery study. Earlier we noted that draft-eligibility is a promising instrument for Vietnam-era veteran status. The draft-eligibility ceilings were RSN 195 for men born in 1950, RSN 125 for men born in 1951, and RSN 95 for men born in 1952. In practice, however, there is a richer link between draft lottery numbers (which we'll call R_i, short for RSN) and veteran status (D_i) than draft-eligibility status alone. Although men with numbers above the eligibility ceiling were not drafted, the ceiling was unknown in advance. Some men therefore volunteered in the hope of serving under better terms and gaining some control over the timing of their service. The pressure to become a draft-induced volunteer was high for men with low lottery numbers, but low for men with high numbers. As a result, there is variation in $P[\text{D}_i = 1|\text{R}_i]$ even for values strictly above or below the draft-eligibility cutoff. For example, men born in 1950 with lottery numbers 200 – 225 were more likely to serve than those with lottery numbers 226 – 250, though ultimately no one in either group was drafted.

The Wald estimator using draft-eligibility as an instrument for men born in 1950 compares the earnings of men with $\text{R}_i < 195$ to the earnings of men with $\text{R}_i > 195$. But the previous discussion suggests the possibility of many more comparisons, for example men with $\text{R}_i \leq 25$ vs. men with $\text{R}_i \in [26 - 50]$; men with $\text{R}_i \in [51 - 75]$ vs. men with $\text{R}_i \in [76 - 100]$, and so on, until these 25-number intervals are exhausted. We might also make the intervals finer, comparing, say, men in 5-number or single-number intervals instead of 25-number intervals. The result of this expansion in the set of comparisons is a set of Wald estimators. These sets are complete in that the intervals partition the support of the underlying instrument, while the individual estimators are linearly independent in the sense that their numerators are linearly independent. Finally, each of these Wald estimators consistently estimates the same causal effect, assumed here to be constant, as long as R_i is independent of potential outcomes and correlated with veteran status (i.e., the Wald denominators are not zero).

[9] An exception is the classical measurement error model, where both the variable to be instrument and the instrument are assumed to be continuous. Here, we have in mind IV scenarios involving omitted variables bias.

The possibility of constructing multiple Wald estimators for the same causal effect naturally raises the question of what to do with all of them. We would like to come up with a single estimate that somehow combines the information in the individual Wald estimates efficiently. As it turns out, the most efficient linear combination of a full set of linearly independent Wald estimates is produced by fitting a line through the group means used to construct these estimates.

The grouped data estimator can be motivated directly as follows. As in (4.1.11), we work with a bivariate constant-effects model, which in this case can be written

$$\text{Y}_i = \alpha + \rho \text{D}_i + \eta_i, \tag{4.1.14}$$

where $\rho = \text{Y}_{1i} - \text{Y}_{0i}$ is the causal effect of interest and $\text{Y}_{0i} = \alpha + \eta_i$. Because R_i was randomly assigned and lottery numbers are assumed to have no effect on earnings other than through veteran status, $E[\eta_i|\text{R}_i] = 0$. It therefore follows that

$$E[\text{Y}_i|\text{R}_i] = \alpha + \rho P[\text{D}_i = 1|\text{R}_i], \tag{4.1.15}$$

since $P[\text{D}_i = 1|\text{R}_i] = E[\text{D}_i|\text{R}_i]$. In other words, the slope of the line connecting average earnings given lottery number with the average probability of service by lottery number is equal to the effect of military service, ρ. This is in spite of the fact that the regression Y_i on D_i—in this case, the difference in means by veteran status—almost certainly differs from ρ since Y_{0i} and D_i are likely to be correlated.

Equation (4.1.15) suggests an estimation strategy based on fitting a line to the sample analog of $E[\text{Y}_i|\text{R}_i]$ and $P[\text{D}_i = 1|\text{R}_i]$. Suppose that R_i takes on values $j = 1, ..., \text{J}$. In principle, j might run from 1 to 365, but in Angrist (1990), lottery-number information was aggregated to 69 five-number intervals, plus a 70th for numbers 346-365. We can therefore think of R_i as running from 1 to 70. Let $\bar{y}_j$ and $\hat{p}_j$ denote estimates of $E[\text{Y}_i|\text{R}_i = j]$ and $P[\text{D}_i = 1|\text{R}_i = j]$, while $\bar{\eta}_j$ denotes the average error in (4.1.14). Because sample moments converge to population moments it follows that OLS estimates of ρ in the grouped equation

$$\bar{y}_j = \alpha + \rho \hat{p}_j + \bar{\eta}_j \tag{4.1.16}$$

are consistent. In practice, however, GLS may be preferable since a grouped equation is heteroskedastic with a known variance structure. The efficient GLS estimator for grouped data in a constant-effects linear model is weighted least squares, weighted by the variance of $\bar{\eta}_j$ (see, e.g., Prais and Aitchison, 1954 or Wooldridge, 2006). Assuming the microdata residual is homoskedastic with variance σ^2_η, this variance is $\frac{\sigma^2_\eta}{n_j}$, where n_j is the group size.

The GLS (or weighted least squares) estimator of ρ in equation (4.1.16) is especially important in this context for two reasons. First, the GLS slope estimate constructed from J grouped observations is an asymptotically efficient linear combination of any full set of J−1 linearly independent Wald estimators

(Angrist, 1991). This can be seen without any mathematics: GLS and any linear combination of pairwise Wald estimators are both linear combinations of the grouped dependent variable. Moreover, GLS is the asymptotically efficient linear estimator for grouped data. Therefore we can conclude that there is no better (i.e., asymptotically more efficient) linear combination of Wald estimators than GLS (again, a maintained assumption here is that ρ is constant). The formula for constructing the GLS estimator from a full set of linearly independent Wald estimators appears in Angrist (1988).

Second, just as each Wald estimator is also an IV estimator, the GLS (weighted least squares) estimator of equation (4.1.16) is also 2SLS. The instruments in this case are a full set of dummies to indicate each lottery-number cell. To see why, define the set of dummy instruments $Z_i \equiv \{r_{ji} = 1[\text{R}_i = j]; j = 1, ...J-1\}$. Now, consider the first stage regression of D_i on Z_i plus a constant. Since this first stage is saturated, the fitted values will be the sample conditional means, $\hat{p}_j$, repeated n_j times for each j. The second stage slope estimate is therefore exactly the same as weighted least squares estimation of the grouped equation, (4.1.16), weighted by the cell size, n_j.

The connection between grouped-data and 2SLS is of both conceptual and practical importance. On the conceptual side, any 2SLS estimator using a set of dummy instruments can be understood as a linear combination of all the Wald estimators generated by these instruments one at a time. The Wald estimator in turn provides a simple framework used later in this chapter to interpret IV estimates in the much more realistic world of heterogeneous potential outcomes.

Although not all instruments are inherently discrete and therefore immediately amenable to a Wald or grouped-data interpretation, many are. Examples include the draft lottery number, quarter of birth, twins, and sibling-sex composition instruments we've already discussed. See also the recent studies by Bennedsen, *et al.,* 2007, and Ananat and Michaels, 2008, both of which use dummies for male first births as instruments. Moreover, instruments that have a continuous flavor can often be fruitfully turned into discrete variables. For example, Angrist, Graddy and Imbens (2000) group continuous weather-based instruments into 3 dummy variables, *stormy*, *mixed*, and *clear*, which they then use to estimate the demand fish. This dummy-variable parameterization seems to capture the main features of the relationship between weather conditions and the price of fish.[10]

On the practical side, the grouped-data equivalent of 2SLS gives us a simple tool that can be used to explain and evaluate any IV strategy. In the case of the draft lottery, for example, the grouped model embodies the assumption that the only reason average earnings vary with lottery numbers is the variation in probability of service across lottery-number groups. If the underlying causal relation is linear with constant effects, then equation (4.1.16) should fit the group means well, something we can assess by inspection and, as discussed in the next section, with the machinery of formal statistical inference.

[10] Continuous instruments recoded as dummies can be seen as providing a parsimonious non-parametric model for the underlying first-stage relation, $E[\text{D}_i|\text{Z}_i]$. In homoskedastic models with constant coefficients, the asymptotically efficient instrument is $E[\text{D}_i|\text{Z}_i]$ (Newey, 1990).

Sometimes labor economists refer to grouped-data plots for discrete instruments as Visual Instrumental Variables (VIV).[11] An example appears in Angrist (1990), reproduced here as Figure 4.1.2. This figure shows the relationship between average earnings in 5-number RSN cells and the probability of service in these cells, for the 1981-84 earnings of white men born 1950-53. The slope of the line through these points is an IV estimate of the earnings loss due to military service, in this case about $2,400, not very different from the Wald estimates discussed earlier but with a lower standard error (in this case, about $800).

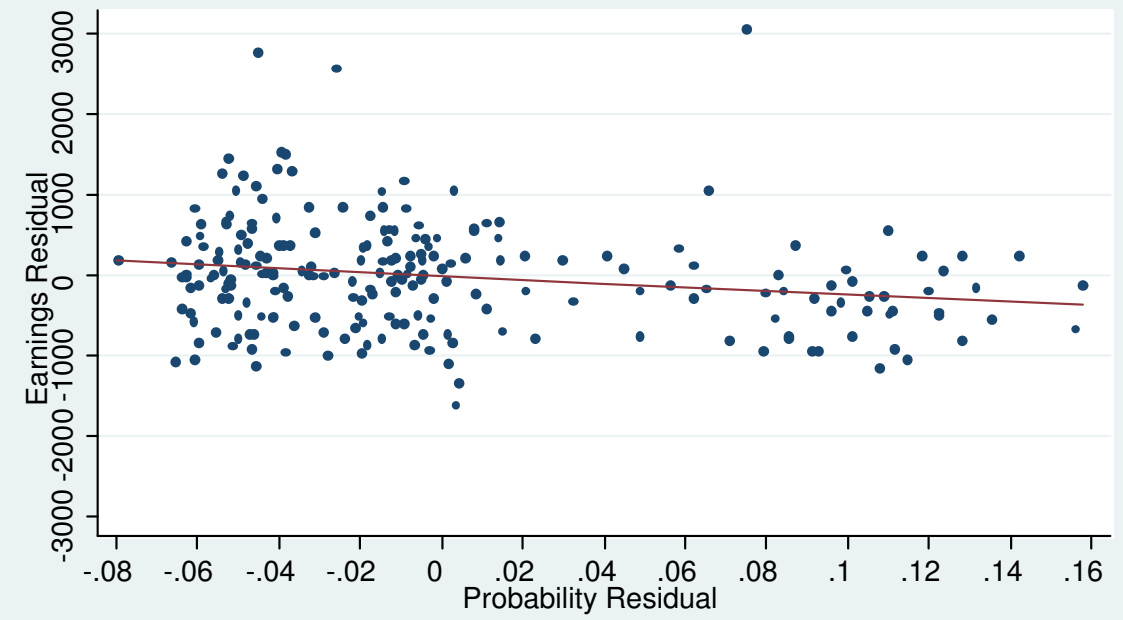

Figure 4.1.2: The relationship between average earnings and the probability of military service (from Angrist 1990). This is a VIV plot of average 1981-84 earnings by cohort and groups of five consecutive draft lottery numbers against conditional probabilities of veteran status in the same cells. The sample includes white men born 1950-53. Plotted points consist of average residuals (over four years of earnings) from regressions on period and cohort effects. The slope of the least-squares regression line drawn through the points is -2,384, with a standard error of 778.

4.2 Asymptotic 2SLS Inference

4.2.1 The Limiting Distribution of the 2SLS Coefficient Vector

We can derive the limiting distribution of the 2SLS coefficient vector using an argument similar to that used in Section 3.1.3 for OLS. In this case, let $V_i \equiv \begin{bmatrix} \mathrm{X}_i' & \hat{s}_i \end{bmatrix}'$ denote the vector of regressors in the 2SLS second

[11] See, e.g., the preface to Borjas (2005).

stage, equation (4.1.9). The 2SLS estimator can then be written

$$\hat{\Gamma}_{2SLS} \equiv \left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i \mathrm{Y}_i,$$

where $\Gamma \equiv \begin{bmatrix} \alpha' & \rho \end{bmatrix}'$ is the corresponding coefficient vector. Note that

$$\begin{aligned} \hat{\Gamma}_{2SLS} &= \Gamma + \left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i[\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)] \\ &= \Gamma + \left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i \eta_i \end{aligned} \tag{4.2.1}$$

where the second equality comes from the fact that the first-stage residuals, $(\mathrm{S}_i - \hat{s}_i)$, are orthogonal to V_i in the sample. The limiting distribution of the 2SLS coefficient vector is therefore the limiting distribution of $\left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i \eta_i$. This quantity is a little harder to work with than the corresponding OLS quantity, because the regressors in this case involve estimated fitted values, $\hat{s}_i$. A Slutsky-type argument shows, however, that we get the same limiting distribution replacing estimated fitted values with the corresponding population fitted values (i.e., replacing $\hat{s}_i$ with $[\mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i]$). It therefore follows that $\hat{\Gamma}_{2SLS}$ has an asymptotically normal distribution, with probability limit Γ, and a covariance matrix estimated consistently by $\left[\sum_i V_i V_i'\right]^{-1} \left[\sum_i V_i V_i' \eta_i^2\right] \left[\sum_i V_i V_i'\right]^{-1}$. This is a sandwich formula like the one for OLS standard errors (White, 1982). As with OLS, if η_i is conditionally homoskedastic given covariates and instruments, the consistent covariance matrix estimator simplifies to $\left[\sum_i V_i V_i'\right]^{-1} \sigma_\eta^2$.

There is little new here, but there is one tricky point. It seems natural to construct 2SLS estimates manually by first estimating the first stage (4.1.4a) and then plugging the fitted values into equation (4.1.9) and estimating this by OLS. That's fine as far as the coefficient estimates go, but the resulting standard errors will be incorrect. Conventional regression software does not know that you are trying to construct a 2SLS estimate. The residual variance estimator that goes into the standard formulas will therefore be incorrect. When constructing standard errors, the software will estimate the residual variance of the equation you estimate by OLS in the second stage:

$$\mathrm{Y}_i - [\alpha'\mathrm{X}_i + \rho\hat{s}_i] = [\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)],$$

replacing the coefficients with the corresponding estimates. The correct residual variance estimator, however, uses the original endogenous regressor to construct residuals and not the first-stage fitted values, $\hat{s}_i$. In other words, the residual you want is $\mathrm{Y}_i - [\alpha'\mathrm{X}_i + \rho\mathrm{S}_i] = \eta_i$, so as to consistently estimate σ_η^2, and not $\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)$. Although this problem is easy to fix (you can construct the appropriate residual variance estimator in a separate calculation), software designed for 2SLS gets this right automatically, and may help

you avoid other common 2SLS mistakes.

4.2.2 Over-identification and the 2SLS Minimand★

Constant-effects models with more instruments than endogenous regressors are said to be over-identified. Because there are more instruments than needed to identify the parameters of interest, these models impose a set of restrictions that can be evaluated as part of a process of specification testing. This process amounts to asking whether the line plotted in a VIV-type picture fits the relevant conditional means tightly enough given the precision with which the means are estimated. The details behind this useful idea are easiest to spell out using matrix notation and a traditional linear model.

Let $Z_i \equiv \left[\begin{array}{cccc} \mathrm{X}_i' & \mathrm{Z}_{1i} & ... & \mathrm{Z}_{\mathrm{Q}i} \end{array} \right]'$ denote the vector formed by concatenating the exogenous covariates and Q instrumental variables and let $W_i \equiv \left[\begin{array}{cc} \mathrm{X}_i' & \mathrm{S}_i \end{array} \right]'$ denote the vector formed by concatenating the covariates and the single endogenous variable of interest. In the quarter-of-birth paper, for example, the covariates are year-of-birth and state-of-birth dummies, the instruments are quarter-of-birth dummies, and the endogenous variable is schooling. The coefficient vector is still $\Gamma \equiv [\alpha', \rho]'$, as in the previous subsection. The residuals for the causal model can be defined as a function of Γ using

$$\eta_i(\Gamma) \equiv \mathrm{Y}_i - \Gamma' W_i = \mathrm{Y}_i - [\alpha' \mathrm{X}_i + \rho \mathrm{S}_i] .$$

This residual is assumed to be uncorrelated with the instrument vector, Z_i. In other words, η_i satisfies the orthogonality condition,

$$E[\mathrm{Z}_i \eta_i(\Gamma)] = 0. \tag{4.2.2}$$

In any sample, however, this equation will not hold exactly because there are more moment conditions than there are elements of Γ.[12] The sample analog of (4.2.2) is the sum over i,

$$\frac{1}{N} \sum Z_i \eta_i(\Gamma) \equiv m_N(\Gamma). \tag{4.2.3}$$

2SLS can be understood as a generalized method of moments (GMM) estimator that chooses a value for Γ by making the sample analog of (4.2.2) as close to zero as possible.

By the central limit theorem, the sample moment vector $\sqrt{N} m_N(\Gamma)$ has an asymptotic covariance matrix equal to $E[Z_i Z_i' \eta_i(\Gamma)^2]$, a matrix we'll call Λ. Although somewhat intimidating at first blush, this is just a matrix of 4th moments, as in the sandwich formula used to construct robust standard errors, (3.1.7). As shown by Hansen (1982), the optimal GMM estimator based on (4.2.2) minimizes a quadratic form in the sample moment vector, $m_N(\hat{g})$, where $\hat{g}$ is a candidate estimator of Γ.[13] The optimal weighting matrix in

[12] With a single endogenous variable and more than one instrument, Γ is $[\mathrm{K}+1] \times 1$, while Z_i is $[\mathrm{K}+\mathrm{Q}] \times 1$ for $\mathrm{Q} > 1$. Hence the resulting linear system cannot be solved unless there is a linear dependency that makes some of the instruments redundant.

[13] "Quadratic form" is matrix language for a weighted sum of squares. Suppose v is an $N \times 1$ vector and M is an $N \times N$

the middle of the GMM quadratic form is Λ^{-1}. In practice, of course, Λ, is unknown and must be estimated. A feasible version of the GMM procedure uses a consistent estimator of Λ in the weighting matrix. Since the estimator using known and estimated Λ have the same limiting distribution, we'll ignore this distinction for now. The quadratic form to be minimized can therefore be written,

$$J_N(\hat{g}) \equiv N m_N(\hat{g})' \Lambda^{-1} m_N(\hat{g}), \tag{4.2.4}$$

where the N-term out front comes from $\sqrt{N}$ normalization of the sample moments. As shown immediately below, when the residuals are conditionally homoskedastic, the minimizer of $J_N(\hat{g})$ is the 2SLS estimator. Without homoskedasticity, the GMM estimator that minimizes (4.2.4) is White's (1982) Two-Stage IV (a generalization of 2SLS) so that it makes sense to call $J_N(\hat{g})$ the "2SLS minimand".

Here are some of the details behind the GMM interpretation of 2SLS[14]. Conditional homoskedasticity means that

$$E[Z_i Z_i' \eta_i (\Gamma)^2] = E[Z_i Z_i'] \sigma_\eta^2.$$

Substituting for Λ^{-1} and using Z,Y and W to denote sample data vectors and matrices, the quadratic form to be minimized becomes

$$J_N(\hat{g}) = (N\sigma_\eta^2)^{-1} \times (\mathrm{Y} - W\hat{g})' Z E[Z_i Z_i']^{-1} Z' (\mathrm{Y} - W\hat{g}). \tag{4.2.5}$$

Finally, substituting the sample cross-product matrix $\left[\frac{Z'Z}{N}\right]$ for $E[Z_i Z_i']$, we have

$$\hat{J}_N(\hat{g}) = (1/\sigma_\eta^2) \times (\mathrm{Y} - W\hat{g})' P_Z (\mathrm{Y} - W\hat{g}),$$

where $P_Z = Z(Z'Z)^{-1}Z$. From here, we get the solution

$$\hat{g} = \hat{\Gamma}_{2SLS} = [W' P_Z W]^{-1} W' P_Z \mathrm{Y}.$$

Since the projection operator, P_Z, produces fitted values, and P_Z is an idempotent matrix, this can be seen to be the OLS estimator of the second-stage equation, (4.1.9), written in matrix notation. More generally, even without homoskedasticity we can obtain a feasible efficient 2SLS-type estimator by minimizing (4.2.4) and using a consistent estimator of $E[Z_i Z_i' \eta_i(\hat{g})^2]$ to form $\hat{J}_N(\hat{g})$. Typically, we'd use the empirical fourth moments, $\sum Z_i Z_i' \hat{\eta}_i^2$, where $\hat{\eta}_i$ is the regular 2SLS residual computed without worrying about heteroskedasticity (see, White, 1982, for distribution theory and other details).

---

matrix. A quadratic form in v is $v'Mv$. If M is a $N \times N$ diagonal matrix with diagonal elements m_i, then $v'Mv = \sum_i m_i v_i^2$.

[14] Much more detailed explanations can be found in Newey (1985), Newey and West (1987), and the original Hansen (1982) GMM paper.

The over-identification test statistic is given by the minimized 2SLS minimand. Intuitively, this statistic tells us whether the sample moment vector, $m_N(\hat{g})$, is close enough to zero for the assumption that $E[Z_i\eta_i] = 0$ to be plausible. In particular, under the null hypothesis that the residuals and instruments are indeed orthogonal, the minimized $J_N(\hat{g})$ has a $\chi^2(\mathrm{Q}-1)$ distribution. We can therefore compare the empirical value of the 2SLS minimand with chi-square tables in a formal testing procedure for $H_0 : E[Z_i\eta_i] = 0$.

For reasons that will soon become apparent, we're not often interested in over-identification *per se*. Our main interest is in the 2SLS minimand when the instruments are a full set of mutually exclusive dummy variables, as for the Wald estimators and grouped-data estimation strategies discussed above. In this important special case, the 2SLS becomes weighted least squares of a grouped equation like (4.1.16), while the 2SLS minimand is the relevant weighted sum of squares being minimized. To see this, note that projection on a full set of mutually exclusive dummy variables for an instrument that takes on J values produces an $N \times 1$ vector of fitted values equal to the J conditional means at each value of the instrument (included covariates are counted as instruments), each one of these n_j times, where n_j is the group size and $\sum n_j = N$. The cross product matrix $[Z'Z]$ in this case is a J×J diagonal matrix with elements n_j. Simplifying, we then have

$$\hat{J}_N(\hat{g}) = (1/\sigma_\eta^2) \times \sum_j n_j(\bar{y}_j - \hat{g}'\bar{W}_j)^2, \tag{4.2.6}$$

where $\bar{W}_j$ is the sample mean of the rows of matrix W in group j. Thus, $\hat{J}_N(\hat{g})$ is the GLS weighted least squares minimand for estimation of the grouped regression: $\bar{y}_j$ on $\bar{W}_j$. With a little bit more work (here we skip the details), we can similarly show that the efficient Two-Step IV procedure without homoskedasticity minimizes

$$\hat{J}_N(\hat{g}) = \sum_j \left(\frac{n_j}{\sigma_j^2}\right)(\bar{y}_j - \hat{g}'\bar{W}_j)^2, \tag{4.2.7}$$

where σ_j^2 is the variance of η_i in group j. Estimation using (4.2.7) is feasible because we can estimate σ_j^2 in a first-step, say, using inefficient-but-still-consistent 2SLS that ignores heteroskedasticity. Efficient two-step IV estimators are constructed in Angrist (1990, 1991).

The GLS structure of the 2SLS minimand allows us to see the over-identification test statistic for dummy instruments as a simple measure of the goodness of fit of the line connecting $\bar{y}_j$ and $\bar{W}_j$. In other words, this is the chi-square goodness of fit statistic for the line in a VIV plot like figure 4.1.2. The chi-square degrees of freedom parameter is given by the difference between the number of values taken on by the instrument and the number of parameters being estimated[15].

Like the various paths leading to the 2SLS estimator, there are many roads to the test-statistic, (4.2.7), as well. Here are two further paths that are worth knowing. First, the test-statistic based on the general GMM minimand for IV, whether the instruments are group dummies or not, is the same as the over-

[15] If, for example, the instrument takes on three values, one of which is assigned to the constant, and the model includes a constant and a single the endogenous variable only, the test statistic has 1 degree of freedom.

identification test statistic discussed in many widely-used econometric references on simultaneous equations models. For example, this statistic features in Hausman's (1983) chapter on simultaneous equations in the *Handbook of Econometrics*, which also proposes a simple computational procedure: for homoskedastic models, the minimized 2SLS minimand is the sample size times the R^2 from a regression of the 2SLS residuals on the instruments (and the included exogenous covariates). The formula for this is $N\left[\frac{\hat{\eta}'P_Z\hat{\eta}}{\hat{\eta}'\hat{\eta}}\right]$, where $\hat{\eta} = \text{Y} - W\hat{\Gamma}_{2SLS}$ is the vector of 2SLS residuals.

Second, it's worth emphasizing that the essence of over-identification can be said to be "more than one way to skin the same econometric cat." In other words, given more than one instrument for the same causal relation, we might consider constructing simple IV estimators one at a time and comparing them. This comparison checks over-identification directly: If each just-identified estimator is consistent, the distance between them should be small relative to sampling variance, and should shrink as the sample size and hence the precision of these estimates increases. In fact, we might consider formally testing whether all possible just-identified estimators are the same. The resulting test statistic is said to generate a Wald[16] test of this null, while the test-statistic based on the 2SLS minimand is said to be a Lagrange Multiplier (LM) test because it can be related to the score vector in a maximum likelihood version of the IV setup.

In the grouped-data version of IV, the Wald test amounts to a test of equality for the set of all possible linearly independent Wald estimators. If, for example, lottery numbers are divided into 4 groups based on various cohorts eligibility cutoffs (RSN 1-95, 96-125, 126-195, and the rest), then 3 linearly independent Wald estimators can be constructed. Alternatively, the efficient grouped-data estimator can be constructed by running GLS on these four conditional means. Four groups means there are 3 possible Wald estimators and 2 non-redundant equality restrictions on these three; hence, the relevant *Wald statistic* has 2 degrees of freedom. On the other hand, 4 groups means three instruments and a constant available to estimate a model with 2 parameters (the constant and the causal effect of military service). So the 2SLS minimand generates an over-identification test statistic with $4 - 2 = 2$ degrees of freedom. And, in fact, provided you use the same method of estimating the weighting matrix in the relevant quadratic forms, these two test statistics not only test the same thing, they are numerically equivalent. This makes sense since we have already seen that 2SLS is the efficient linear combination of Wald estimators.[17]

Finally, a caveat regarding over-identification tests in practice: In our experience, the "over-ID statistic" is often of little value in applied work. Because $J_N(\hat{g})$ measures variance-normalized goodness of-fit, the over-ID test-statistic tends to be low when the underlying estimates are imprecise. Since IV estimates are very often imprecise, we cannot take much satisfaction from the fact that one estimate is within sampling variance of another even if the individual estimates appear precise enough to be informative. On the other

[16] The Wald estimator and Wald test are named after the same statistician, Abraham Wald, but the latter reference is Wald (1943).

[17] The fact that Wald and LM testing procedures for the same null are equivalent in linear models was established by Newey and West (1987). Angrist (1991) gives a formal statement of the argument in this paragraph.

hand, in cases where the underlying IV estimates are quite precise, the fact that the over-ID statistic rejects need not point to an identification failure. Rather, this may be evidence of treatment effect heterogeneity, a possibility we discuss further below. On the conceptual side, however, an understanding of the anatomy of the 2SLS minimand is invaluable, for it once again highlights the important link between grouped data and IV. This link takes the mystery out of estimation and testing with instrumental variables and forces us to confront the raw moments that are the foundation for causal inference.

4.3 Two-Sample IV and Split-Sample IV★

The GMM interpretation of 2SLS highlights the fact that the IV estimator can be constructed from sample moments alone, with no micro data. Returning to the sample moment condition, (4.2.3), and re-arranging slightly produces a regression-like equation involving second moments:

$$\frac{Z'\mathrm{Y}}{N} = \frac{Z'W}{N}\Gamma + \frac{Z'\eta}{N} \tag{4.3.1}$$

GLS estimates of Γ in (4.3.1) are consistent because $E\left[\frac{Z'\mathrm{Y}}{N}\right] = E\left[\frac{Z'W}{N}\right]\Gamma$.

The 2SLS minimand can be thought of as GLS applied to equation (4.3.1), after multiplying by $\sqrt{N}$ to keep the residual from disappearing as the sample size gets large. In other words, 2SLS minimizes a quadratic form in the residuals from (4.3.1) with a (possibly non-diagonal) weighting matrix.[18] An important insight that comes from writing the 2SLS problem in this way is that we do not need the individual observations in our sample to estimate (4.3.1). Just as with the OLS coefficient vector, which can be constructed from the sample conditional mean function, IV estimators can also be constructed from sample moments. The moments needed for IV are $\frac{Z'\mathrm{Y}}{N}$ and $\frac{Z'W}{N}$. The dependent variable, $\frac{Z'\mathrm{Y}}{N}$, is a vector of dimension $[\mathrm{K}+\mathrm{Q}] \times 1$. The regressor matrix, $\frac{Z'W}{N}$, is of dimension $[\mathrm{K}+\mathrm{Q}] \times [\mathrm{K}+1]$. The second-moment equation cannot be solved exactly unless $\mathrm{Q}= 1$ so it makes sense to make the fit as good as possible by minimizing a quadratic form in the residuals. The most efficient weighting matrix for this purpose is the asymptotic covariance matrix of $\frac{Z'\eta}{\sqrt{N}}$. This again produces the 2SLS minimand, $\hat{J}_N(\hat{g})$.

A related insight is the fact that the moment matrices on the left and right hand side of the equals sign in equation (4.3.1) need not come from the same data sets provided these data sets are drawn from the same population. This observation leads to the two-sample instrumental variables (TSIV) estimator used by Angrist (1990) and developed formally in Angrist and Krueger (1992)[19]. Briefly, let Z_1 and Y_1 denote

[18] A quadratic form is the matrix-weighted product, $x'Ax$, where x is a random vector of, say, dimension K and A is a $\mathrm{K}\times\mathrm{K}$ matrix of constants.

[19] Applications of TSIV include Bjorklund and Jantti (1997), Jappelli, Pischke, and Souleles (1998), Currie and Yelowitz (2000), and Dee and Evans (2003). In a recent paper, Inoue and Solon (2005) compare the asymptotic distributions of alternative TSIV estimators, and introduce a maximum likelihood (LIML-type) version of TSIV. They also correct a mistake in the distribution theory in Angrist and Krueger (1995), discussed further, below.

the instrument/covariate matrix and dependent variable vector in data set 1 of size N_1 and let Z_2 and W_2 denote the instrument /covariate matrix and endogenous variable/covariate matrix in data set 2 of size N_2. Assuming $plim\left(\frac{Z_2'W_2}{N_2}\right) = plim\left(\frac{Z_1'W_1}{N_1}\right)$, GLS estimates of the two-sample moment equation

$$\frac{Z_1'\mathrm{Y}_1}{N_1} = \frac{Z_2'W_2}{N_2}\Gamma + \left\{\left[\frac{Z_1'W_1}{N_1} - \frac{Z_2'W_2}{N_2}\right]\Gamma + \frac{Z_1'\eta_1}{N_1}\right\}$$

are also consistent for Γ. The limiting distribution of this estimator is obtained by normalizing by $\sqrt{N_1}$ and assuming $plim\left(\frac{N_2}{N_1}\right)$ is a constant.

The utility of TSIV comes from the fact that it widens the scope for IV estimation to situations where observations on dependent variables, instruments, and the endogenous variable of interest are hard to find in a single sample. It may be easier to find one data set that has information on outcomes and instruments, with which the reduced form can be estimated, and another data set which has information on endogenous variables and instruments, with which the first stage can be estimated. For example, in Angrist (1990), administrative records from the Social Security Administration (SSA) provide information on the dependent variable (annual earnings) and the instruments (draft lottery numbers coded from dates of birth, as well as covariates for race and year of birth). The SSA, however, does not track participants' veteran status. This information was taken from military records, which also contain dates of birth that can used to code lottery numbers. Angrist (1990) used these military records to construct $\frac{Z_2'W_2}{N_2}$, the first-stage correlation between lottery numbers and veteran status conditional on race and year of birth, while the SSA data were used to construct $\frac{Z_1'\mathrm{Y}_1}{N_1}$.

Two further simplifications make TSIV especially easy to use. First, as noted previously, when the instruments consist of a full set of mutually exclusive dummy variables, as in Angrist (1990) and Angrist and Krueger (1992), the second moment equation, (4.3.1), simplifies to a model for conditional means. In particular, the 2SLS minimand for the two-sample problem becomes

$$\hat{J}_N(\hat{g}) = \sum_j \omega_j \left(\bar{y}_{1j} - \hat{g}'\bar{W}_{2j}\right)^2, \tag{4.3.2}$$

where $\bar{y}_{1j}$ is the mean of the dependent variable at instrument/covariate value j in one sample, $\bar{W}_{2j}$ is the mean of endogenous variables and covariates at instrument/covariate value j in a second sample, and ω_j is an appropriate weight. This amounts to weighted least squares estimation of the VIV equation, except that the dependent and independent variables do not come from the same sample. Again, Angrist (1990) and Angrist and Krueger (1992) provide illustrations. The optimal weights for asymptotically efficient TSIV are given by variance of $\bar{y}_{1j} - \hat{g}'\bar{W}_{2j}$. This variance is affected by the fact that moments come from different samples, as are the TSIV standard errors, which are easy to compute in the dummy-instrument case since the estimator is equivalent to weighted least squares.

Second, Angrist and Krueger (1995) introduced a computationally attractive TSIV-type estimator that requires no matrix manipulation and can be implemented with ordinary regression software. This estimator, called Split-Sample IV (SSIV), works as follows.[20] The first-stage estimates in data set two are given by $(Z_2'Z_2)^{-1}Z_2'W_2$. These fitted values can be carried over to data set 1 by constructing the *cross-sample fitted value*, $\hat{W}_{12} \equiv Z_1(Z_2'Z_2)^{-1}Z_2'W_2$. The SSIV second stage is a regression of Y_1 on $\hat{W}_{12}$. The correct limiting distribution for this estimator is derived in Inoue and Solon (2005), who show that the limiting distribution presented in Angrist and Krueger (1992) requires the assumption that $Z_1'Z_1 = Z_2'Z_2$ (as would be true if the marginal distribution of the instruments and covariates is fixed in repeated samples). It's worth noting, however, that the limiting distributions of SSIV and 2SLS are the same when the coefficient on the endogenous variable is zero. The standard errors for this special case are simple to construct and probably provide a reasonably good approximation to the general case.[21]

4.4 IV with Heterogeneous Potential Outcomes

The discussion of IV up to this point postulates a constant causal effect. In the case of a dummy variable like veteran status, this means $\mathrm{Y}_{1i}-\mathrm{Y}_{0i} = \rho$ for all i, while with a multi-valued treatment like schooling, this means $Y_{si} - Y_{s-1,i} = \rho$ for all s and all i. Both are highly stylized views of the world, especially the multi-valued case which imposes linearity as well as homogeneity. To focus on one thing at a time in a heterogeneous-effects model, we start with a zero-one causal variable. In this context, we'd like to allow for treatment-effect heterogeneity, in other words, a distribution of causal effects across individuals.

Why is treatment-effect heterogeneity important? The answer lies in the distinction between the two types of validity that characterize a research design. *Internal validity* is the question of whether a given design successfully uncovers causal effects for the population being studied. A randomized clinical trial or, for that matter, a good IV study, has a strong claim to internal validity. *External validity* is the predictive value of the study's findings in a different context. For example, if the study population in a randomized trial is especially likely to benefit from treatment, the resulting estimates may have little external validity. Likewise,

[20] Angrist and Krueger called this estimator SSIV because they were concerned with a scenario where a single data set is deliberately split in two. As discussed in Section (4.6.4), the resulting estimator may have less bias than conventional 2SLS. Inoue and Solon (2005) refer to the estimator Angrist and Krueger (1995) called SSIV as Two-sample 2SLS or TS2SLS.

[21] This shortcut formula uses the standard errors from the manual SSIV second stage. The correct asymptotic covariance matrix formula, from Inoue and Solon (2005), is

$$\{B[(\sigma_{11} + \kappa\Gamma'\Sigma_{22}\Gamma)A]^{-1}B\}^{-1}$$

where B=$plim\left(\frac{Z_2'W_2}{N_2}\right) = plim\left(\frac{Z_1'W_1}{N_1}\right)$, $A = plim\left(\frac{Z_1'Z_1}{N_1}\right) = plim\left(\frac{Z_2Z_2}{N_2}\right)$, $plim\left(\frac{N_2}{N_1}\right) = \kappa$, σ_{11} is the variance of the reduced-form residual in data set 1, and Σ_{22} is the variance of the first-stage residual in data set 2. In principle, these pieces are easy enough to calculate. Other approaches to SSIV inference include those of Dee and Evans (2003), who calculate standard errors for just-identified models using the delta-method, and Bjorklund and Jantti (1997), who use a bootstrap.

draft-lottery estimates of the effects of conscription for service in the Vietnam era need not be a good measure of the consequences of voluntary military service. An econometric framework with heterogeneous treatment effects helps us to assess both the internal and external validity of IV estimates.[22]

4.4.1 Local Average Treatment Effects

In an IV framework, the engine that drives causal inference is the instrument, Z_i, but the variable of interest is still D_i. This feature of the IV setup leads us to adopt a generalized potential-outcomes concept, indexed against both instruments and treatment status. Let $\text{Y}_i(d, z)$ denote the potential outcome of individual i were this person to have treatment status $\text{D}_i = d$ and instrument value $\text{Z}_i = z$. This tells us, for example, what the earnings of i would be given alternative combinations of veteran status and draft-eligibility status. The causal effect of veteran status given i's realized draft-eligibility status is $\text{Y}_i(1,\text{Z}_i)-\text{Y}_i(0,\text{Z}_i)$, while the causal effect of draft-eligibility status given i's veteran status is $\text{Y}_i(\text{D}_i, 1)-\text{Y}_i(\text{D}_i, 0)$.

We can think of instrumental variables as initiating a causal chain where the instrument, Z_i, affects the variable of interest, D_i, which in turn affects outcomes, Y_i. To make this precise, we need notation to express the idea that the instrument has a causal effect on D_i. Let D_{1i} be i's treatment status when $\text{Z}_i = 1$, while D_{0i} is i's treatment status when $\text{Z}_i = 0$. Observed treatment status is therefore

$$\text{D}_i = \text{D}_{0i} + (\text{D}_{1i} - \text{D}_{0i})\text{Z}_i = \pi_0 + \pi_{1i}\text{Z}_i + \xi_i. \tag{4.4.1}$$

In random-coefficients notation, $\pi_0 \equiv E[\text{D}_{0i}]$ and $\pi_{1i} \equiv (\text{D}_{1i}-\text{D}_{0i})$, so π_{1i} is the heterogeneous causal effect of the instrument on D_i. As with potential outcomes, only one of the potential treatment assignments, D_{1i} and D_{0i}, is ever observed for any one person. In the draft lottery example, D_{0i} tells us whether i would serve in the military if he draws a high (draft-ineligible) lottery number, while D_{1i} tells us whether i would serve if he draws a low (draft-eligible) lottery number. We get to see one or the other of these potential assignments depending on Z_i. The average causal effect of Z_i on D_i is $E[\pi_{1i}]$.

The first assumption in the heterogeneous framework is that the instrument is as good as randomly assigned: it is independent of the vector of potential outcomes and potential treatment assignments. Formally, this can be written

$$[\{\text{Y}_i(d, z); \forall\ d, z\}, \text{D}_{1i}, \text{D}_{0i}] \amalg \text{Z}_i, \tag{4.4.2}$$

Independence is sufficient for a causal interpretation of the *reduced form*, i.e., the regression of Y_i on Z_i.

[22] The distinction between internal and external validity is relatively new to applied econometrics but has a long history in social science. See, for example, the chapter-length discussion in Shadish, Cook, and Campbell (2002), the successor to a classic text on research methods by Campbell and Stanley (1963).

Specifically,

$$\begin{aligned} E\left[\mathrm{Y}_i|\mathrm{Z}_i=1\right]-E\left[\mathrm{Y}_i|\mathrm{Z}_i=0\right] &= E\left[\mathrm{Y}_i(\mathrm{D}_{1i},1)|\mathrm{Z}_i=1\right]-E\left[\mathrm{Y}_i(\mathrm{D}_{0i},0)|\mathrm{Z}_i=0\right] \\ &= E\left[\mathrm{Y}_i(\mathrm{D}_{1i},1)-\mathrm{Y}_i(\mathrm{D}_{0i},0)\right], \end{aligned}$$

the causal effect of the instrument on Y_i. Independence also means that

$$\begin{aligned} E\left[\mathrm{D}_i|\mathrm{Z}_i=1\right]-E\left[\mathrm{D}_i|\mathrm{Z}_i=0\right] &= E\left[\mathrm{D}_{1i}|\mathrm{Z}_i=1\right]-E\left[\mathrm{D}_{0i}|\mathrm{Z}_i=0\right] \\ &= E\left[\mathrm{D}_{1i}-\mathrm{D}_{0i}\right], \end{aligned}$$

in other words, the first-stage from our earlier discussion of 2SLS captures the causal effect of Z_i on D_i.

The second key assumption in the heterogeneous-outcomes framework is the presumption that $\mathrm{Y}_i(d,z)$ is only a function of d.[23] To be specific, while draft-eligibility clearly affects veteran status, an individual's potential earnings *as a veteran* are assumed to be unchanged by draft-eligibility status; while potential earnings *as a nonveteran* are similarly unaffected. In general, the claim that an instrument operates through a single known causal channel is called an *exclusion restriction*. In a linear model with constant effects, the exclusion restriction is expressed by the omission of the instrument from the causal equation of interest, or, equivalently, $E[\mathrm{Z}_i\eta_i]=0$ in equation (4.1.14). It's worth noting that the traditional error-term notation used for simultaneous equations models doesn't lend itself to a clear distinction between independence and exclusion. We need Z_i and η_i to be uncorrelated in this equation, but the reasoning that lies behind this assumption is unclear until we consider both the independence and exclusion restrictions.

The exclusion restriction fails for draft-lottery instruments if men with low draft lottery numbers were affected in some way other than through an increased likelihood of service. For example, Angrist and Krueger (1992) looked for an association between draft lottery numbers and schooling. Their idea was that educational draft deferments would have led men with low lottery numbers to stay in college longer than they would have otherwise desired. If so, draft lottery numbers are correlated with earnings for at least two reasons: an increased likelihood of military service and an increased likelihood of college attendance. The fact that the lottery number is randomly assigned (and therefore satisfies the independence assumption) does not make this possibility less likely. The exclusion restriction is distinct from the claim that the instrument is (as good as) randomly assigned. Rather, it is a claim about a unique channel for causal effects of the instrument.[24]

Using the exclusion restriction, we can define potential outcomes indexed solely against treatment status

[23] Hirano, Imbens, Rubin and Zhou (2000) note that the exclusion restriction that $\mathrm{Y}_i(d,z)$ *equals* $\mathrm{Y}_i(d,z')$ can be weakened to require only that the distributions of $\mathrm{Y}_i(d,z)$ and $\mathrm{Y}_i(d,z')$ be the same.

[24] As it turns out, there is not much of a relationship between schooling and lottery numbers in the Angrist and Krueger (1992) data, probably because educational deferments were phased out during the lottery period.

using the single-index (Y_{1i},Y_{0i}) notation we have been using all along. In particular,

$$\begin{aligned} Y_{1i} &\equiv Y_i(1,1) = Y_i(1,0); \\ Y_{0i} &\equiv Y_i(0,1) = Y_i(0,0). \end{aligned} \tag{4.4.3}$$

The observed outcome, Y_i, can therefore be written in terms of potential outcomes as:

$$\begin{aligned} Y_i &= Y_i(0,Z_i) + [Y_i(1,Z_i) - Y_i(0,Z_i)]D_i \\ &= Y_{0i} + (Y_{1i} - Y_{0i})D_i. \end{aligned} \tag{4.4.4}$$

A random-coefficients notation for this is

$$Y_i = \alpha_0 + \rho_i D_i + \eta_i,$$

a compact version of (4.4.4) with $\alpha_0 \equiv E[Y_{0i}]$ and $\rho_i \equiv Y_{1i} - Y_{0i}$.

A final assumption needed for heterogeneous IV models is that either $\pi_{1i} \geq 0$ for all i or $\pi_{1i} \leq 0$ for all i. This *monotonicity* assumption, introduced by Imbens and Angrist (1994), means that while the instrument may have no effect on some people, all of those who are affected are affected in the same way. In other words, either $D_{1i} \geq D_{0i}$ or $D_{1i} \leq D_{0i}$ for all i. In what follows, we assume monotonicity holds with $D_{1i} \geq D_{0i}$. In the draft-lottery example, this means that although draft-eligibility may have had no effect on the probability of military service for some men, there is no one who was actually kept out of the military by being draft-eligible. Without monotonicity, instrumental variables estimators are not guaranteed to estimate a weighted average of the underlying individual causal effects, $Y_{1i} - Y_{0i}$.

Given the exclusion restriction, the independence of instruments and potential outcomes, the existence of a first stage, and monotonicity, the Wald estimand can be interpreted as the effect of veteran status on those whose treatment status can be changed by the instrument. This parameter is called the local average treatment effect ((LATE); Imbens and Angrist, 1994). Here is a formal statement:

Theorem 4.4.1 *THE LATE THEOREM. Suppose*

(A1, Independence) $\{Y_i(D_{1i},1), Y_{0i}(D_{0i},0), D_{1i}, D_{0i}\} \amalg Z_i$*;*

(A2, Exclusion) $Y_i(d,0) = Y_i(d,1) \equiv Y_{di}$ *for* $d = 0, 1$*;*

(A3, First-stage), $E[D_{1i} - D_{0i}] \neq 0$

(A4, Monotonicity) $D_{1i} - D_{0i} \geq 0 \forall i$*, or vice versa;*

Then

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}] = E[\rho_i|\pi_{1i} > 0].$$

Proof. Use the exclusion restriction to write $E[Y_i|Z_i = 1] = E[Y_{0i} + (Y_{1i} - Y_{0i})D_i|Z_i = 1]$, which equals

$E[\text{Y}_{0i} + (\text{Y}_{1i} - \text{Y}_{0i})\text{D}_{1i}]$ by independence. Likewise $E[\text{Y}_i|\text{Z}_i = 0] = E[\text{Y}_{0i} + (\text{Y}_{1i} - \text{Y}_{0i})\text{D}_{0i}]$, so the numerator of the Wald estimator is $E[(\text{Y}_{1i} - \text{Y}_{0i})(\text{D}_{1i} - \text{D}_{0i})]$. Monotonicity means $\text{D}_{1i} - \text{D}_{0i}$ equals one or zero, so

$$E[(\text{Y}_{1i} - \text{Y}_{0i})(\text{D}_{1i} - \text{D}_{0i})] = E[\text{Y}_{1i} - \text{Y}_{0i}|\text{D}_{1i} > \text{D}_{0i}]P[\text{D}_{1i} > \text{D}_{0i}].$$

A similar argument shows

$$E[\text{D}_i|\text{Z}_i = 1] - E[\text{D}_i|\text{Z}_i = 0] = E[\text{D}_{1i} - \text{D}_{0i}] = P[\text{D}_{1i} > \text{D}_{0i}].$$

■

This theorem says that an instrument which is as good as randomly assigned, affects the outcome through a single known channel, has a first-stage, and affects the causal channel of interest only in one direction, can be used to estimate the average causal effect on the affected group. Thus, IV estimates of effects of military service using the draft lottery estimate the effect of military service on men who served because they were draft-eligible, but would not otherwise have served. This obviously excludes volunteers and men who were exempted from military service for medical reasons, but it includes men for whom draft policy was binding.

How useful is LATE? No theorem answers this question, but it's always worth discussing. Part of the interest in the effects of Vietnam-era service revolves around the question of whether veterans (especially, conscripts) were adequately compensated for their service. Internally valid draft lottery estimates answer this question. Draft lottery estimates of the effects of Vietnam-era conscription may also be relevant for discussions of any future conscription policy. On the other hand, while draft lottery instruments produce internally valid estimates of the causal effect of Vietnam-era conscription, the external validity - i.e., the predictive value of these estimates for military service in other times and places - is not directly addressed by the IV framework. There is nothing in IV formulas to explain *why* Vietnam-era service affects earnings; for that, you need a theory.[25]

You might wonder why we need monotonicity for the LATE theorem, an assumption that plays no role in the traditional simultaneous-equations framework with constant effects. A failure of monotonicity means the instrument pushes some people into treatment while pushing others out. Angrist, Imbens, and Rubin (1996) call the latter group *defiers*. Defiers complicate the link between LATE and the reduced form. To see why, go back to the step in the proof of the LATE theorem which shows the reduced form is

$$E[\text{Y}_i|\text{Z}_i = 1] - E[\text{Y}_i|\text{Z}_i = 0] = E[(\text{Y}_{1i} - \text{Y}_{0i})(\text{D}_{1i} - \text{D}_{0i})].$$

[25] Angrist (1990) interprets draft lottery estimates as the penalty for lost labor market experience. This suggests draft lottery estimates should have external validity for the effects of conscription in other periods, a conjecture born out by the results for WWII draftees in Angrist and Krueger (1994).

Without monotonicity, this is equal to

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_{1i} > \mathrm{D}_{0i}]P[\mathrm{D}_{1i} > \mathrm{D}_{0i}] - E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_{1i} < \mathrm{D}_{0i}]P[\mathrm{D}_{1i} < \mathrm{D}_{0i}].$$

We might therefore have a scenario where treatment effects are positive for everyone yet the reduced form is zero because effects on compliers are canceled out by effects on defiers. This doesn't come up in a constant-effects model because the reduced form is always the constant effect times the first stage regardless of whether the first stage includes defiant behavior.[26]

A deeper understanding of LATE can be had by linking it to a workhorse of contemporary econometrics, the latent-index model for "dummy endogenous variables" like assignment to treatment. These models describe individual choices as determined by a comparison of partly observed and partly unknown ("latent") utilities and costs (see, e.g., Heckman, 1978). Typically, these unobservables are thought of as being related to outcomes, in which case the treatment variable is said to be endogenous (though it is not really endogenous in a simultanenous-equations sense). For example (ignoring covariates), we might model veteran status as

$$\mathrm{D}_i = \begin{cases} 1 & \text{if } \gamma_0 + \gamma_1 \mathrm{Z}_i > v_i \\ 0 & \text{otherwise} \end{cases},$$

where v_i is a random factor involving unobserved costs and benefits of military service assumed to be independent of Z_i. This latent-index model characterizes potential treatment assignments as:

$$\mathrm{D}_{0i} = 1[\gamma_0 > v_i] \text{ and } \mathrm{D}_{1i} = 1[\gamma_0 + \gamma_1 > v_i].$$

Note that in this model, monotonicity is automatically satisfied since γ_1 is a constant. Assuming $\gamma_1 > 0$, LATE can be written

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_{1i} > \mathrm{D}_{0i}] = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\gamma_0 + \gamma_1 > v_i > \gamma_0],$$

which is a function of the latent first-stage parameters, γ_0 and γ_1, as well as the joint distribution of $\mathrm{Y}_{1i} - \mathrm{Y}_{0i}$ and v_i. This is not, in general, the same as the population average treatment effect, $E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}]$, or the

[26] With a constant effect, ρ,

$$\begin{aligned} E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_{1i} \quad &> \quad \mathrm{D}_{0i}]P[\mathrm{D}_{1i} > \mathrm{D}_{0i}] \\ -E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_{1i} \quad &< \quad \mathrm{D}_{0i}]P[\mathrm{D}_{1i} < \mathrm{D}_{0i}]. \\ &= \quad \rho\{P[\mathrm{D}_{1i} > \mathrm{D}_{0i}] - P[\mathrm{D}_{1i} < \mathrm{D}_{0i}]\} \\ &= \quad \rho\{E[\mathrm{D}_{1i} - \mathrm{D}_{0i}]\}. \end{aligned}$$

So a zero reduced form effect means either the first stage is zero or $\rho = 0$.

effect on the treated, $E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_i=1]$. We explore the distinction between different average causal effects in Section 4.4.2.

4.4.2 The Compliant Subpopulation

The LATE framework partitions any population with an instrument into a set of three instrument-dependent subgroups, defined by the manner in which members of the population react to the instrument:

Definition 4.4.1 *Compliers. The subpopulation with* $\mathrm{D}_{1i}=1$ *and* $\mathrm{D}_{0i}=0$.

Always-takers. The subpopulation with $\mathrm{D}_{1i}=\mathrm{D}_{0i}=1$.

Never-takers. The subpopulation with $\mathrm{D}_{1i}=\mathrm{D}_{0i}=0$.

LATE is the effect of treatment on the population of compliers. The term "compliers" comes from an analogy with randomized trials where some experimental subjects comply with the randomly assigned treatment protocol (e.g., take their medicine) but some do not, while some control subjects obtain access to the experimental treatment even though they were not supposed to. Those who don't take their medicine when randomly assigned to do so are never-takers while those who take the medicine even when put into the control group are always-takers. Without adding further assumptions (e.g., constant causal effects), LATE is not informative about effects on never-takers and always-takers because, by definition, treatment status for these two groups is unchanged by the instrument (random assignment). The analogy between IV and a randomized trial with partial compliance is more than allegorical - IV solves the problem of causal inference in a randomized trial with partial compliance. This important point merits a separate subsection, below.

Before turning to this important special case, we make a few general points. First, the average causal effect on compliers is not usually the same as the average treatment effect on the treated. From the simple fact that $\mathrm{D}_i=\mathrm{D}_{0i}+(\mathrm{D}_{1i}-\mathrm{D}_{0i})\mathrm{Z}_i$, we learn that the treated population consists of two non-overlapping groups. By monotonicity, we cannot have both $\mathrm{D}_{0i}=1$ and $\mathrm{D}_{1i}-\mathrm{D}_{0i}=1$ since $\mathrm{D}_{0i}=1$ implies $\mathrm{D}_{1i}=1$. The treated therefore have *either* $\mathrm{D}_{0i}=1$ or $\mathrm{D}_{1i}-\mathrm{D}_{0i}=1$ and $\mathrm{Z}_i=1$, and hence D_i can be written as the sum of two mutually-exclusive dummies, D_{i0} and $(\mathrm{D}_{1i}-\mathrm{D}_{0i})\mathrm{Z}_i$. The treated consist of either always-takers or compliers with the instrument switched on. Since the instrument is as good as randomly assigned, compliers with the instrument switched on are representative of all compliers. From here we get

$$
\begin{aligned}
&\underbrace{E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_i=1\right]}_{\text{effect on the treated}} \qquad (4.4.5)\\
&\quad= E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{0i}=1]P[\mathrm{D}_{0i}=1|\mathrm{D}_i=1]\\
&\qquad +E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{1i}>\mathrm{D}_{0i},\mathrm{Z}_i=1\right]P\left[\mathrm{D}_{1i}>\mathrm{D}_{0i},\mathrm{Z}_i=1|\mathrm{D}_i=1\right]\\
&\quad= \underbrace{E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{0i}=1]}_{\text{effect on always-takers}}P[\mathrm{D}_{0i}=1|\mathrm{D}_i=1]\\
&\qquad +\underbrace{E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{1i}>\mathrm{D}_{0i}\right]}_{\text{effect on compliers}}P\left[\mathrm{D}_{1i}>\mathrm{D}_{0i},\mathrm{Z}_i=1|\mathrm{D}_i=1\right]
\end{aligned}
$$

Since $P[\text{D}_{0i}=1|\text{D}_i=1]$ and $P[\text{D}_{1i}>\text{D}_{0i},\text{Z}_i=1|\text{D}_i=1]$ add up to one, this means that the effect of treatment on the treated is a weighted average of effects on always-takers and compliers.

Likewise, LATE is not the average causal effect of treatment on the non-treated, $E[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_i=0]$. In the draft-lottery example, the average effect on the non-treated is the average causal effect of military service on the population of non-veterans from the Vietnam-era cohorts. The average effect of treatment on the non-treated is a weighted average of effects on never-takers and compliers. In particular,

$$\begin{aligned} & \underbrace{E\left[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_i=0\right]}_{\text{effect on the non-treated}} \\ = \; & \underbrace{E\left[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_{1i}=0\right]}_{\text{effect on never-takers}} P\left[\text{D}_{1i}=0|\text{D}_i=0\right] \\ & + \underbrace{E\left[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_{1i}>\text{D}_{0i}\right]}_{\text{effect on compliers}} P\left[\text{D}_{1i}>\text{D}_{0i},\text{Z}_i=0|\text{D}_i=0\right], \end{aligned} \tag{4.4.6}$$

where we use the fact that, by monotonicity, those with $\text{D}_{1i}=0$ must be never-takers.

Finally, averaging (4.4.5) and (4.4.6) using

$$E[\text{Y}_{1i}-\text{Y}_{0i}]=E[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_i=1]P[\text{D}_i=1]+E[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_i=0]P[\text{D}_i=0]$$

shows the overall population average treatment effect to be a weighted average of effects on compliers, always-takers, and never-takers. Of course, this is a conclusion we could have reached directly given monotonicity and the definition at the beginning of this subsection.

Because an instrumental variable is not directly informative about effects on always-takers and never-takers, instruments do not usually capture the average causal effect on all of the treated or on all of the non-treated. There are important exceptions to this rule, however: instrumental variables that allow no always-takers or no never-takers. Although this scenario is not typical, it is an important special case. One example is the twins instrument for fertility, used by Rosenzweig and Wolpin (1980), Bronars and Grogger (1994), Angrist and Evans (1998), and Angrist, Lavy, and Schlosser (2006). Another is Oreopoulos' (2006) recent study using changes in compulsory attendance laws as instruments for schooling in Britain.

To see how this special case works with twins instruments, let T_i be a dummy variable indicating multiple second births. Angrist and Evans (1998) used this instrument to estimate the causal effect of having three children on earnings in the population of women with at least two children. The third child is especially interesting because reduced fertility for American wives in the 1960s and 1970s meant a switch from three children to two. Multiple second births provide quasi-experimental variation on this margin. Let Y_{0i} denote potential earnings if a woman has only two children while Y_{1i} denotes her potential earnings if she has three, an event indicated by D_i. Assuming that T_i is randomly assigned, i.e., that fertility increases by at most one child in response to a multiple birth, and that multiple births affect outcomes only by increasing fertility,

LATE using the twins instrument, T_i, is also $E[Y_{1i}-Y_{0i}|D_i = 0]$, the average causal effect on women who are not treated (i.e., have two children only). This is because all women who have a multiple second birth end up with three children, i.e., there are no never-takers in response to the twins instrument.

Oreopoulos (2006) also uses IV to estimate an average causal effect of treatment on the non-treated. His study estimates the economic returns to schooling using an increase in the British compulsory attendance age from 14 to 15. Compliance with the Britain's new compulsory attendance law was near perfect, though many teens would previously have dropped out of school at age 14. The causal effect of interest in this case is the earnings premium for an additional year of high-school. Finishing this year can be thought of as the treatment. Since everybody in Oreopoulos' British sample finishes the additional year when compulsory schooling laws are made stricter, Oreopoulos' IV strategy captures the average causal effect of obtaining one more year of high school on all those who leave school at 14. This turns on the fact that British teens are remarkably law-abiding people - Oreopoulos' IV strategy wouldn't estimate the effect of treatment on the non-treated in, say, Israel, where teenagers get more leeway when it comes to compulsory school attendance. Israeli econometricians using changes in compulsory attendance laws as instruments must therefore make do with LATE.

4.4.3 IV in Randomized Trials

The language of the LATE framework is based on an *analogy* between IV and randomized trials. But some instruments really come from randomized trials. If the instrument is a randomly assigned offer of treatment, then LATE is the effect of treatment on those who comply with the offer but are not treated otherwise. An especially important case is when the instrument is generated by a randomized trial with one-sided non-compliance. In many randomized trials, participation is voluntary among those randomly assigned to receive treatment. On the other hand, no one in the control group has access to the experimental intervention. Since the group that receives (i.e., complies with) the assigned treatment is a self-selected subset of those offered treatment, a comparison between those actually treated and the control group is misleading. The selection bias in this case is almost always positive: those who take their medicine in a randomized trial tend to be healthier; those who take advantage of randomly assigned economic interventions like training programs tend to earn more anyway.

IV using the randomly assigned treatment intended as an instrumental variable for treatment received solves this sort of compliance problem. Moreover, LATE is the effect of treatment on the treated in this case. Suppose the instrument, Z_i, is a dummy variable indicating random assignment to a treatment group, while D_i is a dummy indicating whether treatment was actually received. In practice, because of non-compliance, D_i is not equal to Z_i. An example is the randomized evaluation of the JTPA training program, where only 60 percent of those assigned to be trained received training, while roughly 2 percent of those assigned to the control group received training anyway (Bloom, *et al.*, 1997). Non-compliance in the JTPA arose from lack

of interest among participants and the failure of program operators to encourage participation. Since the compliance problem in this case is largely confined to the treatment group, LATE using random assignment, Z_i, as an instrument for treatment received, D_i, is the effect of treatment on the treated.

This use of IV to solve the compliance problems is illustrated in Table 4.4.1, which presents results from the JTPA experiment. The outcome variable of primary interest in the JTPA experiment is total earnings in the 30-month period after random assignment. Columns 1-2 of the table show the difference in earnings between those who were trained and those who were not (the estimates in column 2 are from a regression model that adjusts for a number of individual characteristics measured at the beginning of the experiment. The contrast reported in columns 1-2 is on the order of $4,000 for men and $2,200 for women, in both cases a large treatment effect that amounts to about 20 percent of average earnings. But these estimates are misleading because they compare individuals according to D_i, the actual treatment received. Since individuals assigned to the treatment group were free to decline (and 40% did so), this comparison throws away the random assignment unless the decision to accept treatment is itself independent of potential outcomes. This seems unlikely.

Table 4.4.1: Results from the JTPA experiment: OLS and IV estimates of training impacts

	Comparisons by Training Status		Comparisons by Assignment Status		Instrumental Variable Estimates	
	Without Covariates	With Covariates	Without Covariates	With Covariates	Without Covariates	With Covariates
	(1)	(2)	(3)	(4)	(5)	(6)
A. Men	3,970	3,754	1,117	970	1,825	1,593
	(555)	(536)	(569)	(546)	(928)	(895)
B. Women	2,133	2,215	1,243	1,139	1,942	1,780
	(345)	(334)	(359)	(341)	(560)	(532)

Notes: The table reports OLS, reduced-form, and IV estimates of the effect of subsidized training on earnings in the JTPA experiment. Columns (1) and (2) show differences in earnings by training status; columns (3) and (4) show differences by random-assignment status. Columns (5) and (6) report the result of using random-assignment status as an instrument for training. The covariates used in columns (2), (5) and (6) are *High school or GED, Black, Hispanic, Married, Worked less than 13 weeks in past year, AFDC* (for women), plus indicators for the service strategy recommended, age group and second follow-up survey. Robust standard errors are shown in parenthesis.

Columns 3 and 4 of Table 4.4.1. compare individuals according to whether they were *offered* treatment. In other words, this comparison is based on randomly assigned Z_i. In the language of clinical trials, the contrast in columns 3-4 is known as the *intention-to-treat (ITT) effect*. The intention-to-treat effects in the table are on the order \$1,200 (somewhat less with covariates). Since Z_i was randomly assigned, the ITT effect have a causal interpretation: they tell us the causal effect of the offer of treatment, building in the fact that many of those offered will decline. For this reason, the ITT effect is too small relative to the average causal effect on those who were in fact treated. Columns 5 and 6 put the pieces together and give us the most interesting effect: intention-to-treat divided by the difference in compliance rates between treatment and control groups as originally assigned (about .6). These figures, roughly \$1,800, estimate the effect of treatment on the treated.

How do we know the that ITT-divided-by-compliance is the effect of treatment on the treated? We can recognize ITT as the reduced-form effect of the randomly assigned offer of treatment, our instrument in this case. The compliance rate is the first stage associated with this instrument, and the Wald estimand, as always, is the reduced-form divided by the first-stage. In general this equals LATE, but because we have (almost) no always-takers, the treated population consists (almost) entirely of compliers. The IV estimates in column 5 and 6 of Table 4.4.1 are therefore consistent estimates of the effect of treatment on the treated.

This conclusion is important enough that it warrants an alternative derivation. To the best of our knowledge the first person to point out that the IV formula can be used to estimate the effect of treatment on the treated in a randomized trial with one-sided non-compliance was Howard Bloom (1984). Here is Bloom's result with a simple direct proof.

Theorem 4.4.2 *THE BLOOM RESULT. Suppose the assumptions of the LATE theorem hold, and* $E[D_i|Z_i = 0] = 0$. *Then*

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1]} = E[Y_{1i} - Y_{0i}|D_i = 1].$$

Proof. $E[Y_i|Z_i = 1] = E[Y_{i0} + (Y_{1i} - Y_{0i})D_i|Z_i = 1]$, while $E[Y_i|Z_i = 0] = E[Y_{i0}|Z_i = 0]$ because $E[D_i|Z_i = 0] = 0$. Therefore

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = E[(Y_{1i} - Y_{0i})D_i|Z_i = 1]$$

by independence. But

$$E[(Y_{1i} - Y_{0i})D_i|Z_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1, Z_i = 1]P[D_i = 1|Z_i = 1]$$

while $E[D_i|Z_i = 0] = 0$ means $D_i = 1$ implies $Z_i = 1$. Hence, $E[Y_{1i} - Y_{0i}|D_i = 1, Z_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1]$

■

In addition to telling us how to analyze randomized trials with non-compliance, the LATE framework

opens the door to cleverly-designed randomized experiments in settings where it's impossible or unethical to compel treatment compliance. A famous example from the field of Criminology is the Minneapolis Domestic Violence Experiment (MDVE). The MDVE was a pioneering effort to determine the best police response to domestic violence (Sherman and Berk, 1984). In general, police use a number of strategies when on a domestic violence call. These include referral to counseling, separation orders, and arrest. A vigorous debate swirls around the question of whether a hard-line response - arrest and at least temporary incarceration - is productive, especially in view of the fact that domestic assault charges are frequently dropped.

As a result of this debate, the city of Minneapolis authorized a randomized trial where the police response to a domestic disturbance was determined in part by random assignment. The research design used randomly shuffled color-coded charge sheets telling the responding officers to arrest some perpetrators while referring others to counseling or separating the parties. In practice, however, the police were free to overrule the random assignment. For example, an especially dangerous or drunk offender was arrested no matter what. As a result, the actual response often deviated from the randomly assigned response, though the two are highly correlated.

Most published analyses of the MDVE data recognize this compliance problem and focus on ITT effects, i.e., an analysis using the original random assignment and not the treatment actually delivered. But the MDVE data can also be used to get the average causal effect on compliers, in this case those who were arrested because they were randomly assigned to be but would not have been arrested otherwise. The MDVE is analyzed in this spirit in Angrist (2006). Because everyone in the MDVE who was assigned to be arrested was in fact arrested, there are no never-takers. This is an interesting twist and the flip-side of the Bloom scenario: here, we have $\text{D}_{1i} = 1$ for everybody. Consequently, LATE is the effect of treatment on the non-treated, i.e.,

$$E[\text{Y}_{1i} - \text{Y}_{0i} | \text{D}_{1i} > \text{D}_{0i}] = E[\text{Y}_{1i} - \text{Y}_{0i} | \text{D}_i = 0],$$

where D_i indicates arrest. The IV estimates using MDVE data show that arrest reduces repeat offenses sharply, in this case, among the subpopulation that was not arrested.[27]

4.4.4 Counting and Characterizing Compliers

We've seen that, except in special cases, each instrumental variable identifies a unique causal parameter, one specific to the subpopulation of compliers for that instrument. Different valid instruments for the same causal relation therefore estimate different things, at least in principle (an important exception being

[27] Another application of IV to data from a randomized trial is Krueger (1999). This study uses randomly assigned class size as an instrument for actual class size with data from the Tennessee STAR experiment. For students in first grade and higher, actual class size differs from randomly assigned class size in the STAR experiment because parents and teachers move students around in years after the experiment began. Krueger 1999 also illustrates 2SLS applied to a model with variable treatment intensity, as discussed in section 4.5.3.

instruments that allow for perfect compliance on one side or the other). Although different IV estimates are "weighted-up" by 2SLS to produce a single average causal effect, over-identification testing of the sort discussed in Section 4.2.2, where multiple instruments are validated according to whether or not they estimate the same thing, is out the window in a fully heterogeneous world.

Differences in compliant sub-populations might explain variability in treatment effects from one instrument to another. We would therefore like to learn as much as we can about the compliers for different instruments. Moreover, if the compliant subpopulation is similar to other populations of interest, the case for extrapolating estimated causal effects to these other populations is stronger. In this spirit, Acemoglu and Angrist (2000) argue that quarter-of-birth instruments and state compulsory attendance laws (the minimum schooling required before leaving school in your state of birth when you were 14) affect essentially the same group of people and for the same reasons. We therefore expect IV estimates of the returns to schooling from these two sets of instruments to be similar. We might also expect the quarter of birth estimates to predict the impact of contemporary proposals to strengthen compulsory attendance laws.

On the other hand, if the compliant subpopulations associated with two or more instruments are very different, yet the IV estimates they generate are similar, we might be prepared to adopt homogeneous effects as a working hypothesis. This revives the over-identification idea, but puts it at the service of external validity.[28] This reasoning is illustrated by the study of the effects of family size on children's education by Angrist, Lavy, and Schlosser (2006). The Angrist, Lavy, and Schlosser study is motivated by the observation that children from larger families typically end up with less education than those from smaller families. A long-standing concern in research on fertility is whether the observed negative correlation between larger families and worse outcomes is causal. As it turns out, IV estimates of the effect of family size using a number of different instruments, each with very different compliant subpopulations, all generate results showing no effect of family size. Angrist, Lavy, and Schlosser (2006) argue that their results point to a common treatment of zero for just about everybody in the Israeli population they study.

We have already seen that the size of a complier group is easy to measure. This is just the Wald first-stage, since, given monotonicity, we have

$$\begin{aligned} P[\mathrm{D}_{1i}>\mathrm{D}_{0i}] &= E[\mathrm{D}_{1i}-\mathrm{D}_{0i}] \\ &= E[\mathrm{D}_{1i}]-E[\mathrm{D}_{0i}] \\ &= E[\mathrm{D}_i|\mathrm{Z}_i=1]-E[\mathrm{D}_i|\mathrm{Z}_i=0]. \end{aligned}$$

We can also tell what proportion of the treated are compliers since, for compliers, treatment status is

[28] In fact, maintaining the hypothesis that all instruments in an over-identified model are valid, the traditional over-identification test statistic becomes a formal test for treatment-effect heterogeneity.

completely determined by Z_i. Start with the definition of conditional probability:

$$\begin{aligned} P[\text{D}_{1i} \quad > \quad \text{D}_{0i} | \text{D}_i{=}1] &= \frac{P[\text{D}_i{=}1 | \text{D}_{1i}{>}\text{D}_{0i}] P[\text{D}_{1i}{>}\text{D}_{0i}]}{P[\text{D}_i{=}1]} \\ &= \quad \frac{P[\text{Z}_i{=}1](E[\text{D}_i|\text{Z}_i{=}1] - E[\text{D}_i|\text{Z}_i{=}0])}{P[\text{D}_i{=}1]}. \end{aligned} \tag{4.4.7}$$

The second equality uses the fact that $P[\text{D}_i{=}1|\text{D}_{1i} > \text{D}_{0i}] = P[\text{Z}_i{=}1|\text{D}_{1i} > \text{D}_{0i}]$ and that $P[\text{Z}_i{=}1|\text{D}_{1i} > \text{D}_{0i}]$ $= P[\text{Z}_i{=}1]$ by Independence. In other words, the proportion of the treated who are compliers is given by the first stage, times the probability the instrument is switched on, divided by the proportion treated.

Formula (4.4.7) is illustrated here by calculating the proportion of veterans who are draft-lottery compliers. The ingredients are reported in Table 4.4.2. For example, for white men born in 1950, the first stage is .159, the probability of draft-eligibility is $\frac{195}{366}$, and the marginal probability of treatment is .267. From these statistics, we compute that the compliant subpopulation is .32 of the veteran population in this group. The proportion of veterans who were draft-lottery compliers falls to 20 percent for non-white men born in 1950. This is not surprising since the draft-lottery first stage is considerably weaker for non-whites. The last column of the table reports the proportion of *nonveterans* who would have served if they had been draft-eligible. This ranges from 3 percent of non-whites to 10 percent of whites, reflecting the fact that most non-veterans were deferred, ineligible, or unqualified for military service.

Table 4.4.2: Probabilities of compliance in instrumental variables studies

Source	Endogenous Variable (D)	Instrument (Z)	Sample	$P[D=1]$	1st Stage, $P[D_1 > D_0]$	$P[Z=1]$	$P[D_1 > D_0 \mid D=1]$	$P[D_1 > D_0 \mid D=0]$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Angrist (1990)	Veteran Status	Draft eligibility	White men born in 1950	0.267	0.159	0.534	0.318	0.101
			Non-white men born in 1950	0.163	0.060	0.534	0.197	0.033
Angrist and Evans (1998)	More than 2 children	Twins at second birth	Married women aged 21-35 with two or more children in 1980	0.381	0.603	0.008	0.013	0.966
		First two children are of the same sex	Married women aged 21-35 with two or more children in 1980	0.381	0.060	0.506	0.080	0.048
Angrist and Krueger (1991)	High school graduate	Third or fourth quarter birth	Men born between 1930 and 1939	0.770	0.016	0.509	0.011	0.034
Acemoglu and Angrist (2000)	High school graduate	State requires 11 or more years of school attendance	White men aged 40-49	0.617	0.037	0.300	0.018	0.068

Notes: The table shows an analysis of the absolute and relative size of the complier population for a number of instrumental variables. The first-stage, reported in column 6, gives the absolute size of the complier group. Columns 8 and 9 show the size of the complier population relative to the treated and untreated populations.

The effect of *compulsory* military service is the parameter of primary interest in the Angrist (1990) study, so the fact that draft-eligibility compliers are a minority of veterans is not really a limitation of this study. Even in the Vietnam era, most soldiers were volunteers, a little-appreciated fact about Vietnam-era veterans. The LATE interpretation of IV estimates using the draft lottery highlights the fact that other identification strategies are needed to estimate effects of military service on volunteers (some of these are implemented in Angrist, 1998).

The remaining rows in Table 4.4.2 document the size of the compliant subpopulation for the twins and sibling-sex composition instruments used by Angrist and Evans (1998) to estimate the effects of childbearing and for the quarter of birth instruments and compulsory attendance laws used by Angrist and Krueger (1991) and Acemoglu and Angrist (2000) to estimates the returns to schooling. In each of these studies, the compliant subpopulation is a small fraction of the treated group. For example, less than 2 percent of those who graduated from high school did so because of compulsory attendance laws or by virtue of having been born in a late quarter.

The question of whether a small compliant subpopulation is a cause for worry is context-specific. In some cases, it seems fair to say, "you get what you need." With many policy interventions, for example, it is a marginal group that is of primary interest, a point emphasized in McClellan's (1994) landmark IV study of the effects of surgery on heart attack patients. McClellan uses the relative distance to cardiac care facilities to construct instruments for whether an elderly heart-attack patient is treated with a surgical intervention. Most patients get the same treatment either way, but for some, the case for major surgery is marginal. In such cases, providers or patients opt for a less invasive strategy if the nearest surgical facility is far away. McClellan finds little benefit from surgical procedures for this marginal group. Similarly, an increase in the compulsory attendance age to age 18 is clearly irrelevant for the vast majority of American high school students, but it will affect a few who would otherwise drop out. IV estimates suggest the economic returns to schooling for this marginal group are substantial.

The last column of Table 4.4.2 illustrates the special feature of twins instruments alluded to at the end of the previous subsection. As before, let $\text{D}_i = 0$ for women with two children in a sample of women with at least two children, while $\text{D}_i = 1$ indicates women who have more than two. Because there are no never-takers in response to the event of a multiple birth, i.e., all mothers who have twins at second birth end up with (at least) three children, the probability of compliance among those with $\text{D}_i = 0$ is virtually one (the table shows an entry of .97). LATE is therefore the effect on the non-treated, $E[\text{Y}_{1i} - \text{Y}_{0i} | \text{D}_i = 0]$, in this case.

Unlike the size of the complier group, information on the *characteristics* of compliers seems like a tall order because the compliers cannot be individually identified. Because we can't see both D_{1i} and D_{0i} for each individual, we can't just list those with $\text{D}_{1i} > \text{D}_{0i}$ and then calculate the distribution of characteristics for this group. Nevertheless, it's easy to describe the distribution of complier characteristics. To simplify,

we focus here on characteristics - like race or degree completion - that can be described by dummy variables. In this case, everything we need to know can be learned from variation in the first stage across covariate groups.

Let x_{1i} be a Bernoulli-distributed characteristic, say a dummy indicating college graduates. Are sex-composition compliers more or less likely to be college graduates than other women with two children? This question is answered by the following calculation:

$$\frac{P[x_{1i}=1|\mathrm{D}_{1i}>\mathrm{D}_{0i}]}{P[x_{1i}=1]}=\frac{P[\mathrm{D}_{1i}>\mathrm{D}_{0i}|x_{1i}=1]}{P[\mathrm{D}_{1i}>\mathrm{D}_{0i}]}=\frac{E[\mathrm{D}_i|\mathrm{Z}_i=1,x_{1i}=1]-E[\mathrm{D}_i|\mathrm{Z}_i=0,x_{1i}=1]}{E[\mathrm{D}_i|\mathrm{Z}_i=1]-E[\mathrm{D}_i|\mathrm{Z}_i=0]}. \tag{4.4.8}$$

In other words, the relative likelihood a complier is a college graduate is given by the ratio of the first stage for college graduates to the overall first stage.[29]

This calculation is illustrated in Table 4.4.3**,** which reports compliers' characteristics ratios for age at first birth, nonwhite race, and degree completion using twins and same-sex instruments. The table was constructed from the Angrist and Evans (1998) 1980 census extract. Twins compliers are much more likely to be over 30 than the average mother in the sample, reflecting the fact that younger women who had a multiple birth were likely to go on to have additional children anyway. Twins compliers are also more educated than the average mother, while sex-composition compliers are less educated. This helps to explain the smaller 2SLS estimates generated by twins instruments (reported here in Table 4.1.4), since Angrist and Evans (1998) show that the labor supply consequences of childbearing decline with mother's schooling.

[29] A general method for constructing the mean or other features of the distribution of covariates for compliers uses Abadie's (2003) kappa-weighting scheme. For example,

$$E[\mathrm{X}_i|\mathrm{D}_{1i}>\mathrm{D}_{0i}]=\frac{E[\kappa_i\mathrm{X}_i]}{E[\kappa_i]},$$

where

$$\kappa_i=1-\frac{\mathrm{D}_i(1-\mathrm{Z}_i)}{1-P(\mathrm{Z}_i=1|\mathrm{X}_i)}-\frac{(1-\mathrm{D}_i)\mathrm{Z}_i}{P(\mathrm{Z}_i=1\mathrm{X}_i)}.$$

This works because the weighting function, κ_i, "finds compliers," in a sense discussed in Section (4.5.2), below.

Table 4.4.3: Complier-characteristics ratios for twins and sex-composition instruments

Variable	$E[x]$	Twins at second birth		First two children are same sex	
		$E[x\|D_1 > D_0]$	$P[x\|D_1 > D_0]/P[X]$	$E[x\|D_1 > D_0]$	$P[x\|D_1 > D_0]/P[X]$
	(1)	(2)	(3)	(6)	(5)
Age 30 or older at first birth	0.00291	0.00404	1.39	0.00233	0.995
			(0.0201)		(0.374)
Black or hispanic	0.125	0.103	0.822	0.102	0.814
			(0.00421)		(0.0775)
High school graduate	0.822	0.861	1.048	0.815	0.998
			(0.000772)		(0.0140)
College graduate	0.132	0.151	1.14	0.0904	0.704
			(0.00376)		(0.0692)

Notes: The table reports an analysis of complier characteristics for twins and sex-composition instruments. The ratios in columns 3 and 5 give the relative likelihood compliers have the characteristic indicated in each row. Data are from the 1980 Census 5% sample, including married mothers age 21-35 with at least two children, as in Angrist and Evans (1998). The sample size is 254,654 for all columns.

4.5 Generalizing LATE

The LATE theorem applies to a stripped-down causal model where a single dummy instrument is used to estimate the impact of a dummy treatment with no covariates. We can generalize this in three important ways: multiple instruments (e.g., a set of quarter-of-birth dummies), models with covariates (e.g., controls for year of birth), and models with variable and continuous treatment intensity (e.g., years of schooling). In all three cases, the IV estimand is a weighted average of causal effects for instrument-specific compliers. The econometric tool remains 2SLS and the interpretation remains fundamentally similar to the basic LATE result, with a few bells and whistles. 2SLS with multiple instruments produces a causal effect that averages IV estimands using the instruments one at a time; 2SLS with covariates produces an average of covariate-specific LATEs; 2SLS with variable or continuous treatment intensity produces a weighted average derivative along the length of a possibly nonlinear causal response function.

4.5.1 LATE with Multiple Instruments

The multiple-instruments extension is easy to see. This is essentially the same as a result we discussed in the grouped-data context. Consider a pair of dummy instruments, Z_{1i} and Z_{2i}. Without loss of generality, assume these dummies are mutually exclusive (if not, then we can work with a mutually exclusive set of three dummies, $\text{Z}_{1i}(1-\text{Z}_{2i})$,$\text{Z}_{2i}(1-\text{Z}_{1i})$, and $\text{Z}_{1i}\text{Z}_{2i}$). The two dummies can be used to construct Wald estimators. Again, without loss of generality assume monotonicity is satisfied for each with a positive first stage (if not, we can recode the dummies so this is true). Both therefore estimate a version of $E[\text{Y}_{1i}-\text{Y}_{0i}|\text{D}_{1i} > \text{D}_{0i}]$, though the population with $\text{D}_{1i} > \text{D}_{0i}$ differs for Z_{1i} and Z_{2i}.

Instead of Wald estimators, we can use Z_{1i} and Z_{2i} together in a 2SLS procedure. Since these two dummies and a constant exhaust the information in the instrument set, this 2SLS procedure is the same as grouped-data estimation using conditional means defined given Z_{1i} and Z_{2i} (whether or not the instruments are correlated). As in Angrist (1991), the resulting grouped-data estimator is a linear combination of the underlying Wald estimators. In other words, it is a linear combination of the instrument-specific LATEs using the instruments one at a time (in fact, it is the efficient linear combination in a traditional homoskedastic linear constant-effects model).

This argument is not quite complete since we haven't shown that the linear combination of LATEs produced by 2SLS is also a weighted average (i.e., the weights are non-negative and sum to one). The relevant weighting formulas appear in Imbens and Angrist (1994) and Angrist and Imbens (1995). The formulas are a little messy, so here we lay out a simple version based on the two-instrument example. The example shows that 2SLS using Z_{1i} and Z_{2i} together is a weighted average of IV estimates using Z_{1i} and Z_{2i} one at a time. Let

$$\rho_j = \frac{Cov(\text{Y}_i, \text{Z}_{ji})}{Cov(\text{D}_i, \text{Z}_{ji})}; j = 1, 2$$

denote the two IV estimands using Z_{1i} and Z_{2i}.

The (population) first stage fitted values for 2SLS are $\hat{\mathrm{D}}_i = \pi_{11}\mathrm{Z}_{1i} + \pi_{12}\mathrm{Z}_{2i}$. By virtue of the IV interpretation of 2SLS, the 2SLS estimand is

$$\begin{aligned}
\rho_{2SLS} &= \frac{Cov(\mathrm{Y}_i, \hat{\mathrm{D}}_i)}{Cov(\mathrm{D}_i, \hat{\mathrm{D}}_i)} = \frac{\pi_{11}Cov(\mathrm{Y}_i, \mathrm{Z}_{1i})}{Cov(\mathrm{D}_i, \hat{\mathrm{D}}_i)} + \frac{\pi_{12}Cov(\mathrm{Y}_i, \mathrm{Z}_{2i})}{Cov(\mathrm{D}_i, \hat{\mathrm{D}}_i)} \\
&= \left[\frac{\pi_{11}Cov(\mathrm{D}_i, \mathrm{Z}_{1i})}{Cov(\mathrm{D}_i, \hat{\mathrm{D}}_i)}\right]\left[\frac{Cov(\mathrm{Y}_i, \mathrm{Z}_{1i})}{Cov(\mathrm{D}_i, \mathrm{Z}_{1i})}\right] + \left[\frac{\pi_{21}Cov(\mathrm{D}_i, \mathrm{Z}_{2i})}{Cov(\mathrm{D}_i, \hat{\mathrm{D}}_i)}\right]\left[\frac{Cov(\mathrm{Y}_i, \mathrm{Z}_{2i})}{Cov(\mathrm{D}_i, \mathrm{Z}_{2i})}\right] \\
&= \psi\rho_1 + (1-\psi)\rho_2,
\end{aligned}$$

where

$$\psi = \frac{\pi_{11}Cov(\mathrm{D}_i, \mathrm{Z}_{1i})}{\pi_{11}Cov(\mathrm{D}_i, \mathrm{Z}_{1i}) + \pi_{21}Cov(\mathrm{D}_i, \mathrm{Z}_{2i})}$$

is a number between zero and one that depends on the relative strength of each instrument in the first stage. Thus, we have shown that 2SLS is a weighted average of causal effects for instrument-specific compliant subpopulations. Suppose, for example, that Z_{1i} denotes twins births and Z_{2i} indicates same-sex sibships in families with two or more children, both instruments for family size as in Angrist and Evans (1998). A multiple second birth increases the likelihood of having a third child by about .6 while a same-sex sibling pair increases the likelihood of a third birth by about .07. When these two instruments are used together, the resulting 2SLS estimates are a weighted average of the Wald estimates produced by using the instruments one at a time.[30]

4.5.2 Covariates in the Heterogeneous-effects Model

You might be wondering where the covariates have gone. After all, covariates played a starring role in our earlier discussion of regression and matching. Yet the LATE theorem does not involve covariates. This stems from the fact that when we see instrumental variables as a type of (natural or man-made) randomized trial, covariates take a back seat. If, after all, the instrument is randomly assigned, it is likely to be independent of covariates. Not all instruments have this property, however. As with covariates in the regression models in the previous chapter, the main reason why covariates are included in causal analyses using instrumental variables is that the conditional independence and exclusion restrictions underlying IV estimation may be more likely to be valid after conditioning on covariates. Even randomly assigned instruments, like draft-eligibility status, may be valid only after conditioning on covariates. In the case of draft-eligibility, older cohorts were more likely to be draft-eligible because the cutoffs were higher. Because there are year-of-birth (or age) differences in earnings, draft-eligibility status is a valid instrument only after conditioning on year of birth.

[30] Using twins instruments alone, the IV estimate of the effect of a third child on female labor force participation is -.084 (s.e.=.017). The corresponding samesex estimate is -.138 (s.e.=.029). Using both instruments produces a 2SLS estimate of -.098 (.015). The 2SLS weight in this case is .74 for twins, .26 for samesex, due to the much stronger twins first stage.

More formally, IV estimation with covariates may be justified by a *conditional* independence assumption

$$\{\mathrm{Y}_{1i}, \mathrm{Y}_{0i}, \mathrm{D}_{1i}, \mathrm{D}_{0i}\} \amalg \mathrm{Z}_i | \mathrm{X}_i \tag{4.5.1}$$

In other words, we think of the instrumental variables as being "as good as randomly assigned," conditional on covariates, X_i (here we are implicitly maintaining the exclusion restriction as well). A second reason for incorporating covariates is that conditioning on covariates may reduce some of the variability in the dependent variable. This leads to more precise 2SLS estimates under constant conditional effects.

The simplest causal model with covariates is the constant-effects model, with functional form restrictions as follows:

$$\begin{aligned} E[\mathrm{Y}_{0i}|\mathrm{X}_i] &= \mathrm{X}_i'\alpha^* \text{ for a } \mathrm{K} \times 1 \text{ vector of coefficients, } \alpha^*; \\ \mathrm{Y}_{1i} - \mathrm{Y}_{0i} &= \rho. \end{aligned}$$

In combination with (4.5.1), this motivates 2SLS estimation of an equation like (4.1.6) as discussed in Section 4.1.

A straightforward generalization of the constant-effects model allows

$$\mathrm{Y}_{1i} - \mathrm{Y}_{0i} = \rho(\mathrm{X}_i),$$

where $\rho(\mathrm{X}_i)$ is a deterministic function of X_i. This model can be estimated by adding interactions between Z_i and X_i to the first stage and (the same) interactions between D_i and X_i to the second stage. There are now multiple endogenous variables and hence multiple first-stage equations. These can be written

$$\begin{aligned} \mathrm{D}_i &= \mathrm{X}_i'\pi_{00} + \pi_{01}\mathrm{Z}_i + \mathrm{Z}_i\mathrm{X}_i'\pi_{02} + \xi_{0i} \\ \mathrm{D}_i\mathrm{X}_i &= \mathrm{X}_i'\pi_{10} + \pi_{11}\mathrm{Z}_i + \mathrm{Z}_i\mathrm{X}_i'\pi_{12} + \xi_{1i} \end{aligned}$$

The second stage equation in this case is

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho_0\mathrm{D}_i + \mathrm{D}_i\mathrm{X}_i'\rho_1 + \eta_i,$$

so $\rho(\mathrm{X}_i) = \rho_0 + \rho_1'\mathrm{X}_i$. Alternately, a nonparametric version of $\rho(\mathrm{X}_i)$ can be estimated by 2SLS in subsamples stratified on X_i.

The heterogeneous-effects model underlying the LATE theorem also allows for identification based on conditional independence as in (4.5.1), though the estimand is a little more complicated. For each value of

X_i, we define covariate- specific LATE,

$$\lambda(X_i) \equiv E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}, X_i].$$

The "saturate and weight" approach to estimation with covariates is spelled out in the following theorem (from Angrist and Imbens, 1995).

Theorem 4.5.1 *SATURATE AND WEIGHT. Suppose the assumptions of the LATE theorem hold conditional on* X_i. *That is,*

(CA1, Independence) $\{Y_i(D_{1i}, 1), Y_{0i}(D_{0i}, 0), D_{1i}, D_{0i}\} \amalg Z_i | X_i$;

(CA2, Exclusion) $P[Y_i(d, 0) = Y_i(d, 1)|X_i] = 1$ *for* $d = 0, 1$;

(CA3, First-stage), $E[D_{1i} - D_{0i}|X_i] \neq 0$

We also assume monotonicity (A4) holds as before. Consider the 2SLS estimand based on the first stage equation

$$D_i = \pi_X + \pi_{1X} Z_i + \xi_{1i} \tag{4.5.3}$$

and the second stage equation

$$Y_i = \alpha_X + \rho_c D_i + \eta_i$$

where π_X *and* α_X *denote saturated models for covariates (a full set of dummies for all values of* X_i*) and* π'_{1X} *denotes a separate first-stage effect of* Z_i *for every value of* X_i. *Then* $\rho_c = E[\omega(X_i)\lambda(X_i)]$ *where*

$$\begin{aligned} \omega(X_i) &= \frac{V\{E[D_i|X_i, Z_i]|X_i\}}{E[V\{E[D_i|X_i, Z_i]|X_i\}]} \\ &= \frac{E\{P[D_i = 1|X_i, Z_i](1 - P[D_i = 1|X_i, Z_i])|X_i\}}{E[E[D_i|X_i, Z_i](1 - P[D_i = 1|X_i, Z_i])]}. \end{aligned} \tag{4.5.4}$$

.

This theorem says that 2SLS with a fully saturated first stage and a saturated model for covariates in the second stage produces a weighted average of covariate-specific LATEs. The weights are proportional to the average conditional variance of the population first-stage fitted value, $E[D_i|X_i, Z_i]$, at each value of X_i.[31] The theorem comes from he fact that the first stage coincides with $E[D_i|X_i, Z_i]$ when (4.5.3) is saturated (i.e., the first-stage regression recovers the CEF).

In practice, we may not want to work with a model with a first-stage parameter for each value of the covariates. First, there is the risk of bias, as we discuss at the end of this chapter, and second, a big pile of

[31] Note that the variability in $E[D_i|X_i, Z_i]$ conditional on X_i comes from Z_i. So the weighting formula gives more weight to covariate values where the instrument creates more variation in fitted values. The first line of the weight formula, (4.5.4), holds for any endogenous variable in a 2SLS setup. The second is a consequence of the fact that here the endogenous variable is a dummy.

individually-imprecise first-stage estimates is not pretty to look at. It seems reasonable to imagine that models with fewer parameters, say a restricted first stage imposing a constant π_{1X}, nevertheless approximates some kind of covariate-averaged LATE. This turns out to be true, but the argument is surprisingly indirect. The vision of 2SLS as providing a MMSE error approximation to an underlying causal relation was developed by Abadie (2003).

The Abadie approach begins by defining the object of interest to be $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}]$, the CEF for Y_i given treatment status and covariates, for compliers. An important feature of this CEF is that when the conditions of the LATE theorem hold conditional on X_i, it has a causal interpretation. In other words, *for compliers*, treatment-control contrasts conditional on X_i are equal to conditional-on-X_i LATEs:

$$\begin{aligned} & E\left[\mathrm{Y}_i|\mathrm{D}_i=1,\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right]-E\left[\mathrm{Y}_i|\mathrm{D}_i=0,\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right] \\ = \;& E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right] \end{aligned}$$

This follows immediately from the fact that, given (4.5.1), potential outcomes are independent of D_i given X_i and $\mathrm{D}_{1i}>\mathrm{D}_{0i}$.[32] The upshot is that we can imagine running a regression of Y_i on D_i and X_i in the complier population. Although this regression might not give us the CEF of interest (unless it is linear or the model is saturated), it will, as always, provide the MMSE approximation to it. So a regression of Y_i on D_i and X_i in the complier population approximates $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}]$ just like OLS approximates $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i]$. Alas, we do not know who the compliers are, so we cannot sample them. Nevertheless, they can be found, in the following sense:

Theorem 4.5.2 *ABADIE KAPPA. Suppose the assumptions of the LATE theorem hold conditional on covariates,* X_i. *Let* $g(\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i)$ *be any measurable function of* $(\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i)$ *with finite expectation. Define*

$$\kappa_i = 1-\frac{\mathrm{D}_i(1-\mathrm{Z}_i)}{1-P(\mathrm{Z}_i=1|\mathrm{X}_i)}-\frac{(1-\mathrm{D}_i)\mathrm{Z}_i}{P(\mathrm{Z}_i=1|\mathrm{X}_i)}.$$

Then

$$E[g(\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i)|\mathrm{D}_{1i}>\mathrm{D}_{0i}]=\frac{E[\kappa_i g(\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i)]}{E[\kappa_i]}.$$

[32] For compliers,

$$\begin{aligned} & P\left[\mathrm{D}_i=1|\{\mathrm{Y}_{1i},\mathrm{Y}_{0i}\},\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right] \\ = \;& P\left[\mathrm{Z}_i=1|\{\mathrm{Y}_{1i},\mathrm{Y}_{0i}\},\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right]. \end{aligned}$$

And by conditional independence,

$$\begin{aligned} & P\left[\mathrm{Z}_i=1|\{\mathrm{Y}_{1i},\mathrm{Y}_{0i}\},\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right] \\ = \;& P\left[\mathrm{Z}_i=1|\mathrm{X}_i,\mathrm{D}_{1i}>\mathrm{D}_{0i}\right]. \end{aligned}$$

This can be proved by direct calculation using the fact that, given the assumptions of the LATE theorem, any expectation is a weighted average of means for always-takers, never-takers, and compliers. By monotonicity, those with $\mathrm{D}_i(1-\mathrm{Z}_i) = 1$ are always-takers because they have $\mathrm{D}_{0i} = 1$, while those with $(1-\mathrm{D}_i)\mathrm{Z}_i = 1$ are never-takers because they have $\mathrm{D}_{1i} = 0$. Hence, the compliers are the left-out group.

The Abadie theorem has a number of important implications; for example, it crops up again in the discussion of quantile treatment effects. Here, we use it to approximate $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i} > \mathrm{D}_{0i}]$ by linear regression. Specifically, let α_a and β_a solve

$$(\alpha_a, \beta_a) = \arg\min_{a,b} E\{(E[\mathrm{Y}_i|\mathrm{D}_i, \mathrm{X}_i, \mathrm{D}_{1i} > \mathrm{D}_{0i}] - a\mathrm{D}_i - \mathrm{X}_i'b)^2|\mathrm{D}_{1i} > \mathrm{D}_{0i}\}.$$

In other words, $\alpha_a\mathrm{D}_i + \mathrm{X}_i'\beta_a$ gives the MMSE approximation to $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i} > \mathrm{D}_{0i}]$, or fits it exactly if it's linear. A consequence of Abadie's theorem is that this approximating function can be obtained by solving

$$(\alpha_a, \beta_a) = \arg\min_{a,b} E\{\kappa_i(\mathrm{Y}_i - a\mathrm{D}_i - \mathrm{X}_i'b)^2\}, \tag{4.5.5}$$

the kappa-weighted least-squares minimand.[33]

Abadie proposes an estimation strategy (and develops distribution theory) for a procedure which involves first-step estimation of κ_i using parametric or semiparametric models for the function, $p(\mathrm{X}_i) = P(\mathrm{Z}_i = 1|\mathrm{X}_i)$. The estimates from the first step are then plugged into the sample analog of (4.5.5) in the second step. Not surprisingly, when the only covariate is a constant, Abadie's procedure simplifies to the Wald estimator. More surprisingly, minimization of (4.5.5) produces the traditional 2SLS estimator as long as a linear model is used for $p(\mathrm{X}_i)$ in the construction of κ_i. In other words, if $P(\mathrm{Z}_i = 1|\mathrm{X}_i) = \mathrm{X}_i'\pi$ is used when constructing an estimate of κ_i, the Abadie estimand is 2SLS. Thus, we can conclude that whenever $p(\mathrm{X}_i)$ can be fit or closely approximated by a linear model, it makes sense to view 2SLS as an approximation to the complier causal response function, $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i} > \mathrm{D}_{0i}]$. On the other hand, α_a is not, in general, the 2SLS estimand and β_a is not, in general, the vector of covariate effects produced by 2SLS. Still, the equivalence to 2SLS for linear $P(\mathrm{Z}_i = 1|\mathrm{X}_i)$ leads us to think that Abadie's method and 2SLS are likely to produce similar estimates in most applications, with the further implication that we can think of 2SLS as approximating $E[\mathrm{Y}_i|\mathrm{D}_i,\mathrm{X}_i,\mathrm{D}_{1i} > \mathrm{D}_{0i}]$.

The Angrist (2001) re-analysis of Angrist and Evans (1998) is an example where estimates based on (4.5.5) are indistinguishable from 2SLS estimates. Using twins instruments to estimate the effect of a third child on female labor supply generates a 2SLS estimate of -.088 (s.e.=.017), while the corresponding Abadie estimate is -.089 (s.e.=.017). Similarly, 2SLS and Abadie estimates of the effect on hours worked

[33] The class of approximating functions needn't be linear. Instead of $a\mathrm{D}_i + \mathrm{X}_i'b$, it might make sense to use a nonlinear function like an exponential (if the dependent variable is non-negative) or probit (if the dependent variable is zero-one). We return to this point at the end of this chapter. As noted in Section (4.4.4), the kappa-weighting sceme can be used to characterize covariate distributions for compliers as well as to estimate outcome distributions.

are identical at -3.55 (s.e.=.617). This is not a strike against Abadie's procedure. Rather, it supports the notion, which we hold dear, that 2SLS approximates the causal relation of interest.[34]

4.5.3 Average Causal Response with Variable Treatment Intensity★

An important difference between the causal effects of a dummy variable and a variable that takes on the values {0, 1, 2, . . .} is that in the first case, there is only one causal effect for any one person, while in the latter there are many: the effect of going from 0 to 1, the effect of going from 1 to 2, and so on. The potential-outcomes notation we used for schooling recognizes this. Here it is again: let

$$Y_{si} \equiv f_i(s),$$

denote the potential (or latent) earnings that person i would receive after obtaining s years of education. Note that the function $f_i(s)$ has an "i" subscript on it while s does not. The function $f_i(s)$ tells us what i would earn for *any* value of schooling, s, and not just for the realized value, S_i. In other words, $f_i(s)$ answers causal "what if" questions for multinomial S_i.

Suppose that S_i takes on values in the set $\{0, 1, ..., \bar{s}\}$. Then there are $\bar{s}$ unit causal effects, $Y_{si} - Y_{s-1,i}$. A linear causal model assumes these are the same for all s and for all i, obviously unrealistic assumptions. But we need not take these assumptions literally. Rather, 2SLS provides a computational device that generates a weighted average of unit causal effects, with a weighting function we can estimate and study, so as to learn where the action is coming from with a particular instrument. This weighting function tells us how the compliers are distributed over the range of S_i. It tells us, for example, that the returns to schooling estimated using quarter of birth or compulsory schooling laws come from shifts in the distribution of high school grades. Other instruments, like the distance instruments used by Card (1995), act elsewhere on the schooling distribution and therefore capture a different sort of return.

To flesh this out, assume that a single binary instrument, Z_i, a dummy for having been born in a state with restrictive compulsory school laws, is to be used to estimate the returns to schooling (as in Acemoglu and Angrist, 2000). Also, let s_{1i} denote the schooling i would get if $z_i = 1$, and let s_{0i} denote the schooling i would get if $z_i = 0$. The theorem below, from Angrist and Imbens (1995), offers an interpretation of the Wald estimand with variable treatment intensity in this case. Note that here we combine the independence and exclusion restrictions by simply stating that potential outcomes indexed by s are independent of the instruments.

Theorem 4.5.3 *AVERAGE CAUSAL RESPONSE. Suppose*

[34] Abadie (2003) gives formulas for standard errors and Alberto Abadie has posted software to compute them. The bootstrap provides a simple alternative, which we used to construct standard errors for the Abadie estimates mentioned in this paragraph.

(ACR1, Independence and Exclusion) $\{Y_{0i}, Y_{1i}, ..., Y_{\bar{s}i}; s_{0i}, s_{1i}\} \amalg \mathrm{Z}_i$*;*

(ACR2, First-stage), $E[s_{1i} - s_{0i}] \neq 0$

(ACR3, Monotonicity) $s_{1i} - s_{0i} \geq 0 \forall i$*, or vice versa; assume the first*

Then

$$\frac{E[\mathrm{Y}_i|\mathrm{Z}_i = 1] - E[\mathrm{Y}_i|\mathrm{Z}_i = 0]}{E[\mathrm{S}_i|\mathrm{Z}_i = 1] - E[\mathrm{S}_i|\mathrm{Z}_i = 0]} = \sum_{s=1}^{\bar{s}} \omega_s E[Y_{si} - Y_{s-1,i}|s_{1i} \geq s > s_{0i}]$$

where

$$\omega_s = \frac{P[s_{1i} \geq s > s_{0i}]}{\sum_{j=1}^{\bar{s}} P[s_{1i} \geq j > s_{0i}]}$$

The weights ω_s *are non-negative and sum to one.*

The average causal response (ACR) theorem says that the Wald estimator with variable treatment intensity is a weighted average of the *unit causal response* along the length of the potentially nonlinear causal relation described by $f_i(s)$. The unit causal response, $E[Y_{si} - Y_{s-1,i}|s_{1i} \geq s > s_{0i}]$, is the average difference in potential outcomes for *compliers at point* s, i.e., individuals driven by the instrument from a treatment intensity less than s to at least s. For example, the quarter of birth instruments used by Angrist and Krueger (1991) push some people from 11^{th} grade to finishing 12^{th} or higher, and others from 10^{th} grade to finishing 11^{th} or higher. The Wald estimator using quarter of birth instruments combines all of these effects into a single average causal response.

The relative size of the group of compliers at point s is $P[s_{1i} \geq s > s_{0i}]$. By monotonicity, this must be non-negative and is given by the difference in the CDF of S_i at point s. To see this, note that

$$\begin{aligned} P\left[s_{1i} \geq s > s_{0i}\right] &= P\left[s_{1i} \geq s\right] - P\left[s_{0i} \geq s\right] \\ &= P\left[s_{0i} < s\right] - P\left[s_{1i} < s\right], \end{aligned}$$

which is non-negative since monotonicity requires $s_{1i} \geq s_{0i}$. Moreover,

$$P[s_{0i} < s] - P[s_{1i} < s] = P[\mathrm{S}_i < s|\mathrm{Z}_i = 0] - P[\mathrm{S}_i < s|\mathrm{Z}_i = 1]$$

by Independence. Finally, note that because the mean of a non-negative random variable is one minus the CDF, we have,

$$\begin{aligned} & E\left[\mathrm{S}_i|\mathrm{Z}_i = 1\right] - E\left[\mathrm{S}_i|\mathrm{Z}_i = 0\right] \\ = & \sum_{j=1}^{\bar{s}} (P\left[\mathrm{S}_i < j|\mathrm{Z}_i = 1\right] - P\left[\mathrm{S}_i < j|\mathrm{Z}_i = 0\right]) = \sum_{j=1}^{\bar{s}} P\left[s_{1i} \geq j > s_{0i}\right] \end{aligned}$$

Thus, the ACR weighting function can be consistently estimated by comparing the CDFs of the endogenous variables (treatment intensity) with the instrument switched off and on. The weighting function is normalized

by the first-stage.

The ACR theorem helps us understand what we are learning from a 2SLS estimate. For example, instrumental variables derived from compulsory attendance and child labor laws capture the causal effect of increases in schooling in the 6-12 grade range, but not from post-secondary schooling. This is illustrated in Figure 4.5.1, taken from Acemoglu and Angrist (2000).

The figure plots differences in the probability that educational attainment is at or exceeds the grade level on the X-axis (i.e., one minus the CDF). The differences are between men exposed to different child labor laws and compulsory schooling laws in the a sample of white men aged 40-49 drawn from the 1960, 1970, and 1980 censuses. The instruments are coded as the number of years of schooling required either to work (Panel A) or leave school (Panel B) in the year the respondent was aged 14. Men exposed to the least restrictive laws are the reference group. Each instrument (e.g., a dummy for 7 years of schooling required before work is allowed) can be used to construct a Wald estimator by making comparisons with the reference group.

Panel A of Figure 4.5.1 shows that men exposed to more restrictive child labor laws were 1-6 percentage points more likely to complete grades 8-12. The intensity of the shift depends on whether the laws required 7, 8, or 9-plus years of schooling before work was allowed. But in all cases, the CDF differences decline at lower grades, and drop off sharply after grade 12. Panel B shows a similar pattern for compulsory attendance laws, though the effects are a little smaller and the action here is at somewhat higher grades, consistent with the fact that compulsory attendance laws are typically binding in higher grades than child labor laws.

Before wrapping up our discussion of LATE generalizations, it's worth noting that most of the elements we have covered work in combination. For example, models with multiple instruments and variable treatment intensity generate a weighted average of the ACR for each instrument. Likewise, the saturate and weight theorem applies to models with variable treatment intensity. On the other hand, we do not yet have an extension of Abadie's Kappa for models with variable treatment intensity. A final important extension is to the scenario where the causal variable of interest is continuous and we can therefore think of the causal response function as having derivatives.

So Long and Thanks for all the Fish

Suppose that as with the schooling problem, we imagine counterfactuals as being generated by an underlying functional relation. In this case, however, the causal variable of interest can take on any non-negative value and the functional relation is assumed to have a derivative. An example where this makes sense is a demand curve, the quantity demanded as a function of price. In particular, let $q_i(p)$ denote the quantity demanded in market i at hypothetical price p. This is a potential outcome, like $f_i(s)$, except that instead of individuals the unit of observation is a time or a location or both. For example, Angrist, Graddy, and Imbens (2000) estimate the elasticity of quantity demanded at the Fulton wholesale fish market in New York City. The

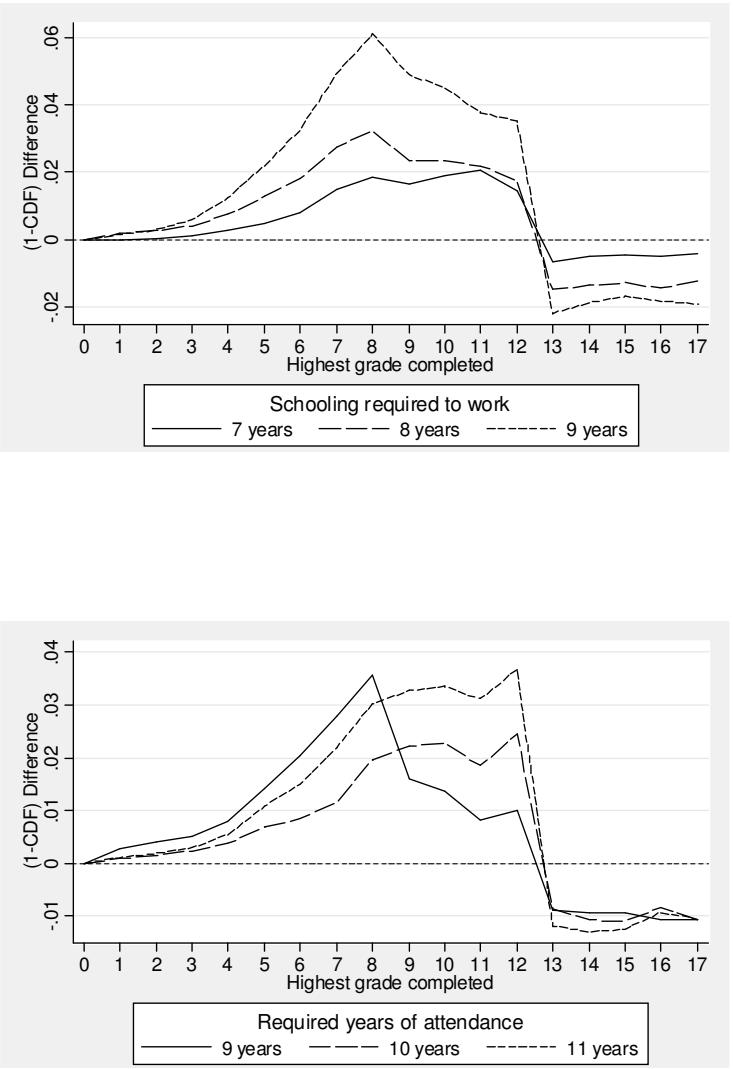

Figure 4.5.1: The effect of compulsory schooling instruments on the probability of schooling (from Acemoglu and Angrist 2000). The figures show the difference in the probability of schooling at or exceeding the grade level on the x-axis. The reference group is 6 or fewer years of required schooling in the top panel, and 8 or fewer years in the bottom panel. The top panel shows the CDF difference by severity of child labor laws. The bottom panel shows the CDF difference by severity of compulsory attendace laws.

slope of this demand curve is $q_i'(p)$; if quantity and price are measured in logs, this is an elasticity.

The instruments in Angrist, Graddy, and Imbens (2000) are derived from data on weather conditions off the coast of Long Island, not too far from major commercial fishing grounds. Stormy weather makes it hard to catch fish, driving up the price, and reducing quantity demanded. Angrist, Graddy, and Imbens use dummy variables such as $stormy_i$, a dummy indicating periods with high wind and waves to estimate the demand for fish. The data consist of daily observations on wholesale purchases of Whiting, a cheap fish used for fish cakes and things like that.

The Wald estimator using the $stormy_i$ instrument can be represented as

$$\frac{E[q_i|stormy_i = 1] - E[q_i|stormy_i = 0]}{E[p_i|stormy_i = 1] - E[p_i|stormy_i = 0]} \tag{4.5.6}$$

$$= \frac{\int E[q_i'(t)|\ p_{1i} \geq t > p_{0i}]P[p_{1i} \geq t > p_{0i}]dt}{\int P[p_{1i} \geq t > p_{0i}]dt}, \tag{4.5.7}$$

where p_i is the price in market (day) i and p_{1i} and p_{0i} are potential prices indexed by $stormy_i$. This is a weighted average derivative with weighting function $P[p_{1i} \geq t > p_{0i}] = P[p_i \leq t|z_i = 0] - P[p_i \leq t|z_i = 1]$ at price t. In other words, IV estimation using $stormy_i$ produces an average of the derivative $q_i'(t)$, with weight given to each possible price (indexed by t) in proportion to the instrument-induced change in the cumulative distribution function (CDF) of prices at that point. This is the same sort of averaging as in the ACR theorem except that now the underlying causal response is a derivative instead of a one-unit difference.

The average causal response formula, (4.5.6), comes from the fact that

$$E[q_i|stormy_i = 1] - E[q_i|stormy_i = 0] = E\int_{p_{0i}}^{p_{1i}} q_i'(t)dt, \tag{4.5.8}$$

by the fundamental theorem of calculus. Two interesting special cases fall neatly out of equation (4.5.8). The first is when the causal response function is linear, i.e., $q_i(p) = \alpha_{0i} + \alpha_{1i}p$, for some random coefficients, α_{0i} and α_{1i}. Then, we have

$$\frac{E[q_i|stormy_i = 1] - E[q_i|stormy_i = 0]}{E[p_i|stormy_i = 1] - E[p_i|stormy_i = 0]} = \frac{E[\alpha_{1i}(p_{1i} - p_{0i})]}{E[p_{1i} - p_{0i}]}, \tag{4.5.9}$$

a weighted average of the random coefficient, α_{1i}. The weights are proportional to the price change induced by the weather in market i.

The second special case is when we can write quantity demanded as

$$q_i(p) = Q(p) + \eta_i, \tag{4.5.10}$$

where $Q(p)$ is a non-stochastic function and η_i is an additive random error. By this we mean $q_i'(p) = Q'(p)$

every day or in every market. In this case, the average causal response function becomes

$$\int Q'(t)\omega(t)dt, \text{where } \omega(t) = \frac{P[p_{1i} \geq t > p_{0i}]}{\int P[p_{1i} \geq r > p_{0i}]dr}.$$

These special cases highlight the two types of averaging wrapped up in the ACR theorem and its continuous corollary, (4.5.6). First, there is averaging *across* markets, with weights proportional to the first-stage impact on prices in each market. Markets where prices are highly sensitive to the weather contribute the most. Second, there is averaging *along* the length of the causal response function in a given market. IV recovers the average derivative over a range of prices where the CDF of prices shifts most sharply.

4.6 IV Details

4.6.1 2SLS Mistakes

2SLS estimates are easy to compute, especially since software like SAS and Stata will do it for you. Occasionally, however, you might be tempted to do it yourself just to see if it really works. Or you may be stranded on the planet Krikkit with all of your software licenses expired (Krikkit is encased in a slo-time envelope, so it will take you a long time to get licenses renewed). "Manual 2SLS" is for just such emergencies. In the Manual 2SLS procedure, you estimate the first stage yourself (which in any case, you should be looking at), and plug the fitted values into the second stage equation, which is then estimated by OLS. Returning to the system at the beginning of this chapter, the first and second stages are

$$\begin{aligned} \mathrm{S}_i &= \mathrm{X}_i'\pi_{10} + \pi_{11}'\mathrm{Z}_i + \xi_{1i} \\ \mathrm{Y}_i &= \alpha'\mathrm{X}_i + \rho\hat{s}_i + [\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)] \end{aligned}$$

where X_i is a set of covariates, Z_i is a set of excluded instruments, and the first stage fitted values are $\hat{s}_i = \mathrm{X}_i'\hat{\pi}_{10} + \pi_{11}'\mathrm{Z}_i$.

Manual 2SLS takes some of the mystery out of canned 2SLS, and may be useful in a software crisis, but it opens the door to mistakes. For one thing, as we discussed earlier, the OLS standard errors from the manual second stage will not be correct (the OLS residual variance is the variance of $\eta_i + \rho(\mathrm{S}_i - \hat{s}_i)$, while for proper 2SLS standard errors you want the variance of η_i only). There are more subtle risks as well.

Covariate Ambivalence

Suppose the covariate vector contains two sorts of variables, some (say, X_{0i}) that you are comfortable with, and others (say, X_{1i}) about which you are ambivalent. Griliches and Mason (1972) faced this scenario when

constructing 2SLS estimates of a wage equation that treats AFQT scores (an ability test used by the armed forces) as an endogenous control variable to be instrumented. The instruments for AFQT are early schooling (completed before military service), race, and family background variables. They estimated a system that can be described like this:

$$\begin{aligned} \text{S}_i &= \text{X}_{0i}'\pi_{10} + \pi_{11}'\text{Z}_i + \xi_{1i} \\ \text{Y}_i &= \alpha_0'\text{X}_{0i} + \alpha_0'\text{X}_{1i} + \rho\hat{s}_i + [\eta_i + \rho(\text{S}_i - \hat{s}_i)]. \end{aligned}$$

This looks a lot like manual 2SLS.

A closer look, however, reveals an important difference between the equations above and the usual 2SLS procedure: the covariates in the first and second stages are not the same. For example, Griliches and Mason included age in the second stage but not in the first, a fact noted by Cardell and Hopkins (1977) in a comment on their paper. This is a mistake. Griliches' and Mason's second stage estimates are not the same as 2SLS. What's worse, they are inconsistent where 2SLS might have been fine. To see why, note that the first-stage residual, $\text{S}_i - \hat{s}_i$, is uncorrelated with X_{0i} *by construction* since OLS residuals are always uncorrelated with included regressors. But because X_{1i} is not included in the first-stage it is likely to be correlated with the first-stage residuals (e.g., age is probably correlated with the AFQT residual from the Griliches and Mason (1972) first stage). The inconsistency from this correlation spills over to all coefficients in the second stage. The moral of the story: put the same exogenous covariates in your first and second stage. If a covariate is good enough for the second stage, it's good enough for the first.

Forbidden Regressions

Forbidden regressions were forbidden by MIT Professor Jerry Hausman in 1975, and while they occasionally resurface in an under-supervised thesis, they are still technically off-limits. A forbidden regression crops up when researchers apply 2SLS reasoning directly to nonlinear models. A common scenario is a dummy endogenous variable. Suppose, for example, the causal model of interest is

$$\text{Y}_i = \alpha'\text{X}_i + \rho\text{D}_i + \eta_i, \tag{4.6.1}$$

where D_i is a dummy variable for veteran status. The usual 2SLS first stage is

$$\text{D}_i = \pi_{10}'\text{X}_i + \pi_{11}'\text{Z}_i + \xi_{1i}, \tag{4.6.2}$$

a linear regression of D_i on covariates and regressors.

Because D_i is a dummy variable, the CEF associated with this first stage, $E[\text{D}_i|\text{X}_i,\text{Z}_i]$, is probably nonlinear. So the usual OLS first-stage is an approximation to the underlying nonlinear CEF. We might,

therefore, use a nonlinear first stage in an attempt to come closer to the CEF. Suppose that we use Probit to model $E[\mathrm{D}_i|\mathrm{X}_i,\mathrm{Z}_i]$. The Probit first stage is $\Phi[\mathrm{X}_i'\pi_{p0} + \pi_{p1}'\mathrm{Z}_i]$, where π_{p0} and π_{p1} are Probit coefficients, and the fitted values are $\hat{\mathrm{D}}_{pi} = \Phi[\mathrm{X}_i'\hat{\pi}_{p0} + \hat{\pi}_{p1}'\mathrm{Z}_i]$. The forbidden regression in this case is the second stage equation created by substituting $\hat{\mathrm{D}}_{pi}$ for D_i:

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \rho\hat{\mathrm{D}}_{pi} + [\eta_i + \rho(\mathrm{D}_i - \hat{\mathrm{D}}_{pi})]. \tag{4.6.3}$$

The problem with (4.6.3) is that only OLS estimation of (4.6.2) is guaranteed to produce first-stage residuals that are uncorrelated with fitted values and covariates. If $E[\mathrm{D}_i|\mathrm{X}_i,\mathrm{Z}_i] = \Phi[\mathrm{X}_i'\pi_{p0} + \pi_{p1}'\mathrm{Z}_i]$, then residuals from the nonlinear model will be asymptotically uncorrelated with X_i and $\hat{\mathrm{D}}_{pi}$, but who is to say that the first stage CEF is really Probit? With garden-variety 2SLS, in contrast, we do not need to worry about whether the first-stage CEF is really linear.[35]

A simple alternative to the forbidden second step, (4.6.3), avoids problems due to an incorrect nonlinear first stage. Instead of plugging in nonlinear fitted values, we can use the nonlinear fitted values *as instruments*. In other words, use $\hat{\mathrm{D}}_{pi}$ as an instrument for (4.6.1) in a conventional 2SLS procedure (as always, the exogenous covariates, X_i, should also be in the instrument list). Use of fitted values as instruments is the same as plugging in fitted values when the first-stage is estimated by OLS, but not in general. Nonlinear-fits-as-instruments has the further advantage that, if the nonlinear model gives a better approximation of the first-stage CEF than the linear model, the resulting 2SLS estimates will be more efficient than those using a linear first stage (Newey, 1990).

But here, too, there is a drawback. The nonlinear-fits-as-instruments procedure implicitly uses nonlinearities in the first stage as a source of identifying information. To see this, suppose the causal model of interest includes the instruments, Z_i :

$$\mathrm{Y}_i = \alpha'\mathrm{X}_i + \gamma'\mathrm{Z}_i + \rho\mathrm{D}_i + \eta_i. \tag{4.6.4}$$

Now, with the first stage given by (4.6.2), the model is unidentified and conventional 2SLS estimates of (4.6.4) don't exist. But 2SLS estimates using X_i, Z_i, $\hat{\mathrm{D}}_{pi}$ do exist, because $\hat{\mathrm{D}}_{pi}$ is a nonlinear function of X_i and Z_i that is excluded from the second stage. Should you use this nonlinearity as a source of identifying information? We usually prefer to avoid this sort of back-door identification since its not clear what the underlying experiment really is.

As a rule, naively plugging in first-stage fitted values in nonlinear models is a bad idea. This includes models with a nonlinear second stage as well as those where the CEF for the first stage is nonlinear. Suppose,

[35] The insight that consistency of 2SLS estimates in a traditional SEM does not depend on correct specification of the first-stage CEF goes back to Kelejian (1971). Use of a nonlinear plug-in first-stage may not do too much damage in practice - a probit first-stage can be pretty close to linear - but why take a chance when you don't have to?

for example, that you believe the causal relation between schooling and earnings is approximately quadratic (as in Card's [1995] structural model). In other words, the model of interest is

$$\mathrm{Y}_i = \alpha' \mathrm{X}_i + \rho_1 \mathrm{S}_i + \rho_2 \mathrm{S}_i^2 + \eta_i. \tag{4.6.5}$$

Given two instruments, it's easy enough to estimate (4.6.5) treating both S_i and S_i^2 as endogenous. In this case, there are two first-stage equations, one for S_i and one for S_i^2. You need at least two instruments for this to work, of course. It's natural to use Z_i and its square (unless Z_i is a dummy, in which case you'll need a better idea).

You might be tempted, however, to work with a single first stage, say equation (4.6.2), and estimate the following second stage manually:

$$\mathrm{Y}_i = \alpha' \mathrm{X}_i + \rho_1 \hat{s}_i + \rho_2 \hat{s}_i^2 + [\eta_i + \rho_1(\mathrm{S}_i - \hat{s}_i) + \rho_2(\mathrm{S}_i^2 - \hat{s}_i^2)].$$

This is a mistake since $\hat{s}_i$ can be correlated with $\mathrm{S}_i^2 - \hat{s}_i^2$ while $\hat{s}_i^2$ can be correlated with both $\mathrm{S}_i - \hat{s}_i$ and $\mathrm{S}_i^2 - \hat{s}_i^2$. On the other hand, as long as X_i and Z_i are uncorrelated with η_i in (4.6.5), and you have enough instruments in Z_i, 2SLS estimation of (4.6.5) is straightforward.

4.6.2 Peer Effects

A vast literature in social science is concerned with peer effects. Loosely speaking, this means the causal effect of group characteristics on individual outcomes. Sometimes regression is used in an attempt to uncover these effects. In practice, the use of regression models to estimate peer effects is fraught with peril. Although this is not really an IV issue *per se*, the language and algebra of 2SLS helps us understand why peer effects are hard to identify.

Broadly speaking, there are two types of peer effects. The first concerns the effect of group characteristics such as the average schooling in a state or city on individually-measured outcome variable. This peer effect links the average of one variable to individual outcomes as described by another variable. For example, Acemoglu and Angrist (2000) ask whether a given individual's earnings are affected by the average schooling in his or her state of residence. The theory of human capital externalities suggests that living in a state with a more educated workforce may make everyone in the state more productive, not just those who are more educated. This kind of spillover is said to be a *social return* to schooling: human capital that benefits everyone, whether or not they are more educated.

A causal model which allows for such externalities can be written

$$Y_{ijt} = \delta_j + \lambda_t + \gamma \overline{S}_{jt} + \rho \mathrm{S}_i + u_{jt} + \eta_{ijt}, \tag{4.6.6}$$

where Y_{ijt} is the log weekly wage of individual i in state j in year t, u_{jt} is a state-year error component, and η_i is an individual error term. The controls δ_j and λ_t are state-of-residence and year effects. The coefficient ρ is the returns to schooling for an individual, while the coefficient γ is meant to capture the effect of average schooling, $\overline{S}_{jt}$, in state j and year t.

In addition to the usual concerns about s_i, the most important identification problem raised by equation (4.6.6) is omitted variables bias from correlation between average schooling and other state-year effects embodied in the error component u_{jt}. For example, public university systems may expand during cyclical upturns, generating a common trend in state average schooling levels and state average earnings. Acemoglu and Angrist (2000) attempt to solve this problem using instrumental variables derived from historical compulsory attendance laws that are correlated with $\overline{S}_{jt}$ but uncorrelated with contemporary u_{jt} and η_i.

While omitted state-year effects are the primary concern motivating Acemoglu and Angrist's (2000) instrumental variables estimation, the fact that one regressor, $\overline{S}_{jt}$, is the average of another regressor, s_i, also complicates the interpretation of OLS estimates of equation (4.6.6). To see this, consider a simpler version of (4.6.6) with a cross-section dimension only. This can be written

$$Y_{ij} = \mu + \pi_0 \mathrm{s}_i + \pi_1 \overline{S}_j + \nu_i; \text{ where } E[\nu_i \mathrm{s}_i] = E[\nu_i \overline{S}_j] \equiv 0. \tag{4.6.7}$$

where Y_{ij} is he log weekly wage of individual i in state j and $\overline{S}_j$ is average schooling in the state. Now, let ρ_0 denote the coefficient from a bivariate regression of Y_{ij} on s_i only and let ρ_1 denote the coefficient from a bivariate regression of Y_{ij} on $\overline{S}_j$ only. From the discussion of grouping and 2SLS earlier in this chapter, it's clear that ρ_1 is the 2SLS estimate of the coefficient on s_i in a bivariate regression of Y_{ij} on s_i using a full set of state dummies as instruments. The Appendix uses this fact to show that the parameters in equation (4.6.7) can be written in terms of ρ_0 and ρ_1 as

$$\begin{aligned} \pi_0 &= \rho_1 + \phi(\rho_0 - \rho_1) \\ \pi_1 &= \phi(\rho_1 - \rho_0) \end{aligned} \tag{4.6.8}$$

where $\phi = \frac{1}{1-R^2} > 1$, and R^2 is the first-stage R-squared.

The upshot of (4.6.8) is that if, *for any reason,* OLS estimates of the bivariate regression of wages on individual schooling differ from 2SLS estimates using state-dummy instruments, the coefficient on average schooling in (4.6.7) will be nonzero. For example, if instrumenting with state dummies corrects for attenuation bias due to measurement error in s_i, we have $\rho_1 > \rho_0$ and the spurious appearance of positive external returns. In contrast, if instrumenting with state dummies eliminates the bias from positive correlation between s_i and unobserved earnings potential, we have $\rho_1 < \rho_0$, and the appearance of negative social returns.[36] In practice, therefore, it is very difficult to substantiate social effects by OLS estimation of an

[36] The coefficient on average schooling in an equation with individual schooling can be interpreted as the Hausman (1978)

equation like 4.6.6, though more sophisticated strategies where both the individual and group averages are treated as endogenous may work.

A second and even more difficult peer effect to uncover is the effect of the group average of a variable on the individual level of *this same variable.* This is not really an IV problem; it takes us back to basic regression issues. To see this point, suppose that $\overline{S}_j$ is the high-school graduation rate in school j, and we would like to know whether students are more likely to graduate from high school when everyone around them is more likely to graduate from high school. To uncover the peer effect in high school graduation rates, we might work with a regression model like:

$$\text{S}_{ij} = \mu + \pi_2 \overline{S}_j + \xi_{ij}, \tag{4.6.9}$$

where S_{ij} is individual i's high school graduation status and $\overline{S}_j$ is the average high school graduation rate in school j, which i attends.

At first blush, equation (4.6.9) seems like a sensible formulation of a well-defined causal question, but in fact it is nonsense. The regression of S_{ij} on $\overline{S}_j$ *always* has a coefficient of 1, a conclusion that can be drawn immediately once you recognize $\overline{S}_j$ as the first-stage fitted value from a regression of S_{ij} on a full set of school dummies.[37] Thus, an equation like (4.6.9) cannot possibly be informative about causal effects.

A modestly improved version of the bad peer regression changes (4.6.9) to

$$\text{S}_{ij} = \mu + \pi_4 \overline{S}_{(i)j} + \xi_{ij}, \tag{4.6.10}$$

where $\overline{S}_{(i)j}$ is the mean of S_{ij} in school j, excluding student i. This is a step in the right direction - by definition, i is not in the group used to construct $\overline{S}_{(i)j}$ - but still problematic because S_{ij} and $\overline{S}_{(i)j}$ are both affected by school-level random shocks. The presence of random effects in the error term raises important issues for statistical inference, issues discussed at length in Chapter 8. But in an equation like (4.6.10), group-level random shocks are more that a problem for standard errors: any shock common to the group (school) creates spurious peer effects. For example, particularly effective school principals may raise graduation rates for everyone in the schools at which they work. This looks like a peer effect since it induces correlation between S_{ij} and $\overline{S}_{(i)j}$ even if there is no causal link between peer means and individual student

test statistic for the equality of OLS estimates and 2SLS estimates of private returns to schooling using state dummies as instruments. Borjas (1992) discusses a similar problem affecting the estimation of ethnic-background effects.

[37] Here is a direct proof that the regression of S_{ij} on $\overline{S}_j$ is always unity:

$$\begin{aligned} \frac{\sum_j \sum_i \text{S}_{ij}(\overline{S}_j - \overline{S})}{\sum_j n_j(\overline{S}_j - \overline{S})^2} &= \frac{\sum_j (\overline{S}_j - \overline{S}) \sum_i \text{S}_{ij}}{\sum_j n_j(\overline{S}_j - \overline{S})^2} \\ &= \frac{\sum_j (\overline{S}_j - \overline{S})(n_j \overline{S}_j)}{\sum_j n_j(\overline{S}_j - \overline{S})^2} = 1. \end{aligned}$$

achievement. We therefore prefer not see regressions like (4.6.10) either.

The best shot at a causal investigation of peer effects focuses on variation in *ex ante* peer characteristics, that is, some measure of peer quality which predates the outcome variable and is therefore unaffected by common shocks. A recent example is Ammermueller and Pischke (2006), who study the link between classmates' family background, as measured by the number of books in their homes, and student achievement in European primary schools. The Ammermueller and Pischke regressions are versions of

$$\text{S}_{ij} = \mu^* + \pi_4 \overline{B}_{(i)j} + \xi_{ij},$$

where $\overline{B}_{(i)j}$ is the average number of books in the home of student i's peers. This looks like (4.6.10), but with an important difference. The variable $\overline{B}_{(i)j}$ is a feature of the home environment that predates test scores and is therefore unaffected by school-level random shocks.

Angrist and Lang (2004) provide another example of an attempt to link student achievement with the *ex ante* characteristics of peers. The Angrist and Lang study looks at the impact of bused-in low-achieving newcomers on high-achieving residents' test scores. The regression of interest in this case is a version of

$$\text{S}_{ij} = \mu + \pi_3 \overline{m}_j + \xi_{ij}, \tag{4.6.11}$$

where $\overline{m}_j$ is the number of bused-in low-achievers in school j and S_{ij} is resident-student i's test score. Spurious correlation due to common shocks is not a concern in this context for two reasons. First, $\overline{m}_j$ is a feature of the school population determined by students outside the sample used to estimate (4.6.11). Second, the number of low-achievers is an *ex ante* variable biased on prior information about where the students come from and not the outcome variable, S_{ij}. School-level random effects remain an important issue for inference, however, since $\overline{m}_j$ is a group-level variable.

4.6.3 Limited Dependent Variables Reprise

In Section 3.4.2, we discussed the consequences of limited dependent variables for regression models. When the dependent variable is binary or non-negative, say, employment status or hours worked, the CEF is typically nonlinear. Most nonlinear LDV models are built around a non-linear transformation of a linear latent index. Examples include Probit, Logit, and Tobit. These models capture features of the associated CEFs (e.g., Probit fitted values are guaranteed to be between zero and one, while Tobit fitted values are non-negative). Yet we saw that the added complexity and extra work required to interpret the results from latent-index models may not be worth the trouble.

An important consideration in favor of OLS is a conceptual robustness that structural models often lack. OLS is always a MMSE linear approximation to the CEF. In fact, we can think of OLS as a scheme for computing marginal effects - a scheme that has the virtue of simplicity, automation, and comparability

across studies. Nonlinear latent-index models are more like GLS - they provide an efficiency gain when taken literally, but they require a commitment to functional form and distributional assumptions about which we do not usually feel strongly.[38] A second consideration is the difference between the latent-index parameters at heart of nonlinear models and the average causal effects that we believe should be the objects of primary interest in most research projects.

The arguments in favor of conventional OLS with LDVs apply with equal force to 2SLS and models with endogenous variables. IV methods capture local average treatment effects regardless of whether the dependent variable is binary, non-negative, or continuously distributed on the real line. With covariates, we can think of 2SLS as estimating LATE averaged across covariate cells. In models with variable or continuous treatment intensity, 2SLS gives us the average causal response or an average derivative. Although Abadie (2003) has shown that 2SLS does not, in general, provide the MMSE approximation to the complier causal response function, in practice, 2SLS estimates come out remarkably close to estimates using the more rigorously grounded Abadie procedure (and with a saturated model for covariates, 2SLS and Abadie are the same). And, of course, 2SLS estimates LATE directly; there is no intermediate step involving the calculation of marginal effects.

2SLS is not the only way to go. An alternative more elaborate approach tries to build up a causal story by describing the process generating LDVs in detail. A good example is bivariate Probit, which can be applied to the Angrist and Evans (1998) example like this. Suppose that a woman decides to have a third child by comparing costs and benefits using a net benefit function or latent index that is linear in covariates and excluded instruments, with a random component or error term, v_i. The bivariate Probit first stage can be written

$$\mathrm{D}_i = 1[\mathrm{X}_i'\gamma_0 + \gamma_1 \mathrm{Z}_i > v_i], \tag{4.6.12}$$

where Z_i is an instrumental variable that increases the benefit of a third child, conditional on covariates, X_i. For example, American parents appear to value a third child more when they have had either two boys or two girls, a sort-of portfolio-diversification phenomenon that can be understood as increasing the benefit

[38] The analogy between nonlinear LDV models and GLS is more than rhetorical. Consider a Probit model with nonlinear CEF $E[\mathrm{Y}_i|\mathrm{X}_i] = \Phi\left[\frac{\mathrm{X}_i'\beta}{\sigma}\right] \equiv p_i$. The first-order conditions for maximum likelihood estimation of this model are

$$\sum \frac{(\mathrm{Y}_i - p_i)\mathrm{X}_i}{p_i(1-p_i)} = 0.$$

Thus, maximum likelihood is the same as GLS estimation of the nonlinear model

$$\mathrm{Y}_i = \Phi\left[\frac{\mathrm{X}_i'\beta}{\sigma}\right] + \xi_i.$$

Consistency of the maximum likelihood estimator turns on the assumption that the conditional variance of Y_i is $p_i(1-p_i)$. It's worth noting that we can dispense with this assumption and simply fit Y_i to $\Phi\left[\frac{\mathrm{X}_i'\beta}{\sigma}\right]$ by nonlinear least squares (NLLS). This sort of agnostic NLLS shares the robustness properties of OLS; it gives the best MMSE fit in a class of approximating functions.

of a third child in families with same-sex sibships.

An outcome of primary interest in this context is employment status, a Bernoulli random variable with a conditional mean between zero and one. To complete the model, suppose that employment status, Y_i, is determined by the latent index

$$\mathrm{Y}_i = 1[\mathrm{X}_i'\beta_0 + \beta_1 \mathrm{D}_i > \varepsilon_i], \tag{4.6.13}$$

where ε_i is a second random component or error term. This latent index can be seen as arising from a comparison of the costs and benefits of working.

The source of omitted variables bias in the bivariate Probit setup is correlation between v_i and ε_i. In other words, unmeasured random determinants of childbearing are correlated with unmeasured random determinants of employment. The model is identified by assuming Z_i is independent of these components, and that the random components are normally distributed. Given normality, the parameters in (4.6.12) and (4.6.13) can be estimated by maximum likelihood. The log likelihood function is

$$\begin{aligned} &\sum \mathrm{Y}_i \ln \Phi_b \left(\frac{\mathrm{X}_i'\beta_0 + \beta_1 \mathrm{D}_i}{\sigma_\varepsilon}, \frac{\mathrm{X}_i'\gamma_0 + \gamma_1 \mathrm{Z}_i}{\sigma_\nu}; \rho_{\varepsilon\nu} \right) \\ &+(1-\mathrm{Y}_i) \ln \left[1 - \Phi_b \left(\frac{\mathrm{X}_i'\beta_0 + \beta_1 \mathrm{D}_i}{\sigma_\varepsilon}, \frac{\mathrm{X}_i'\gamma_0 + \gamma_1 \mathrm{Z}_i}{\sigma_\nu}; \rho_{\varepsilon\nu} \right) \right], \end{aligned} \tag{4.6.14}$$

where $\Phi_b(\cdot, \cdot; \rho_{\varepsilon\nu})$ is the bivariate normal distribution function with correlation coefficient $\rho_{\varepsilon\nu}$. Note, however, that we can multiply the latent index coefficients by a positive constant without changing the likelihood. The object of estimation is therefore the ratio of the index coefficients to the standard deviation of the error terms (e.g., $\beta_1/\sigma_\varepsilon$).

The potential outcomes defined by the bivariate Probit model are

$$\mathrm{Y}_{0i} = 1[\mathrm{X}_i'\beta_0 > \varepsilon_i] \text{ and } \mathrm{Y}_{1i} = 1[\mathrm{X}_i'\beta_0 + \beta_1 > \varepsilon_i],$$

while potential treatment assignments are

$$\mathrm{D}_{0i} = 1[\mathrm{X}_i'\gamma_0 > v_i] \text{ and } \mathrm{D}_{1i} = 1[\mathrm{X}_i'\gamma_0 + \gamma_1 > v_i].$$

As usual, only one potential outcome and one potential assignment is observed for any one person. It's also clear from this representation that correlation between v_i and ε_i is the same thing as correlation between potential treatment assignments and potential outcomes.

The latent index coefficients do not themselves tell us anything about the size of the causal effect of childbearing on employment other than the sign. To see this, note that the average causal effect of childbearing is

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}] = E\{1[\mathrm{X}_i'\beta_0 + \beta_1 > \varepsilon_i] - 1[\mathrm{X}_i'\beta_0 > \varepsilon_i]\}$$

while the average effect on the treated is

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1] = E\{1[\mathrm{X}_i'\beta_0 + \beta_1 > \varepsilon_i] - 1[\mathrm{X}_i'\beta_0 > \varepsilon_i]|\mathrm{X}_i'\gamma_0 + \gamma_1\mathrm{Z}_i > v_i\}.$$

Given alterative distributional assumptions for v_i and ε_i, these can be anything (If the error terms are heteroskedastic then even the sign is indeterminate).

Under normality, the average causal effects generated by the bivariate Probit model are easy to evaluate. The average causal effect is

$$\begin{aligned} & E\left\{1\left[\mathrm{X}_i'\beta_0 + \beta_1 > \varepsilon_i\right] - 1\left[\mathrm{X}_i'\beta_0 > \varepsilon_i\right]\right\} \\ = \; & E\left\{\Phi\left[\frac{\mathrm{X}_i'\beta_0 + \beta_1}{\sigma}\right] - \Phi\left[\frac{\mathrm{X}_i'\beta_0}{\sigma}\right]\right\}, \end{aligned} \tag{4.6.15}$$

where $\Phi[\cdot]$ is the normal CDF. The effect on the treated is a little more complicated since it involves the bivariate normal CDF

$$\begin{aligned} & E\left[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{D}_i = 1\right] \\ = \; & E\left\{\frac{\Phi_b\left(\frac{\mathrm{X}_i'\beta_0+\beta_1}{\sigma_\varepsilon}, \frac{\mathrm{X}_i'\gamma_0+\gamma_1\mathrm{Z}_i}{\sigma_\nu}; \rho_{\varepsilon\nu}\right) - \Phi_b\left(\frac{\mathrm{X}_i'\beta_0}{\sigma_\varepsilon}, \frac{\mathrm{X}_i'\gamma_0+\gamma_1\mathrm{Z}_i}{\sigma_\nu}; \rho_{\varepsilon\nu}\right)}{\Phi\left(\frac{\mathrm{X}_i'\gamma_0+\gamma_1\mathrm{Z}_i}{\sigma_\nu}\right)}.\right\} \end{aligned} \tag{4.6.16}$$

Since the bivariate normal CDF is a canned function in many software packages, this is easy enough to calculate in practice.

Bivariate Probit probably qualifies as harmless in the sense that it's not very complicated, and easy to get right using packaged software routines. Still, it shares the disadvantages of nonlinear latent-index modeling discussed in the previous chapter. First, some researchers become distracted by an effort to identify index coefficients instead of average causal effects. For example, a large literature in econometrics is concerned with the identification of index coefficients without the need for distributional assumptions. Applied researchers interested in causal effects can safely ignore this work.[39]

A second vice in this context is also a virtue. Bivariate Probit and other models of this sort can be used to identify population average causal effects and/or effects on the treated. 2SLS does not promise you average causal effects, only *local* average causal effects. But it should be clear from (4.6.15) that the assumed normality of the latent index error terms is essential for this. As always, the best you can do without a distributional assumption is LATE, the average causal effect for compliers. For bivariate Probit, we can

[39] Suppose the latent error term has an unknown distribution, with CDF $\Lambda[\cdot]$. The average causal effect in this case is

$$E\left\{\Lambda\left[\mathrm{X}_i'\beta_0 + \beta_1\right] - \Lambda\left[\mathrm{X}_i'\beta_0\right]\right\} = \Lambda'[\mathrm{X}_i'\beta_0 + \tilde{\beta}_1]\beta_1,$$

where $\tilde{\beta}_1$ is in $[0, \beta_1]$. This always depends on the shape of $\Lambda[\cdot]$.

write LATE as

$$\begin{aligned} & E\left[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{1i}>\mathrm{D}_{0i}\right] \\ = \; & E\{1[\mathrm{X}_i'\beta_0+\beta_1>\varepsilon_i]-1[\mathrm{X}_i'\beta_0>\varepsilon_i]|\mathrm{X}_i'\gamma_0+\gamma_1>v_i>\mathrm{X}_i'\gamma_0\}, \end{aligned}$$

which, like (4.6.16), can be evaluated using joint normality of v_i and ε_i. But you needn't bother using normality to evaluate $E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_{1i}>\mathrm{D}_{0i}]$, since LATE can be estimated by IV for each X_i and averaged using the histogram of the covariates. Alternately, do 2SLS and settle for a variance-weighted average of covariate-specific LATEs.

You might be wondering whether LATE is enough. Perhaps you would like to estimate the average treatment effect or the effect of treatment on the treated and are willing to make a few extra assumptions to do so. That's all well and good, but in our experience, you can't get blood from a stone, even with heroic assumptions. Since local information is all that's in the data, in practice the average causal effects produced by bivariate Probit are likely to be similar to 2SLS estimates provided the model for covariates is sufficiently flexible. This is illustrated in Table 4.6.1**,** which reports 2SLS and bivariate Probit estimates of the effects of a third child on female labor supply using the Angrist-Evans (1998) same-sex instruments and the same 1980 census sample of married women with 2 or more children used in their paper**.** The dependent variable is a dummy for having worked the previous year; the endogenous variable is a dummy for having a third child. The first stage effect of a same-sex sibship on the probability of a third birth is about 7 percentage points.

Panel A of Table 4.6.1 reports estimates from a model with no covariates. The 2SLS estimate of -.138 in column 1 is numerically identical to the Abadie causal effect estimated using a linear model in column 2, as it should be in this case. Without covariates, the 2SLS slope coefficient provides the best linear approximation to the complier causal response function as does Abadie's kappa-weighting procedure. The marginal effect changes little if, instead of a linear approximation, we use nonlinear least squares with a Probit CEF. The marginal effect estimated by minimizing

$$E\left\{\kappa_i\left(\mathrm{Y}_i-\Phi\left[\frac{\beta_0+\beta_1\mathrm{D}_i}{\sigma_\varepsilon}\right]\right)^2\right\}$$

is -.137, reported in column 3. This is not surprising since the model without covariates imposes no functional form assumptions.

Perhaps more surprising is the fact that marginal effects and the average treatment effects calculated using (4.6.15) and (4.6.16) are also the same as the 2SLS and Abadie estimates. These results are reported in columns 4-6. The marginal effect calculated using a derivative to approximate to the finite difference in (4.6.15) is -.138 (in column 4, labelled MFX for marginal effects), while both average treatment effects are -.139 in columns 5 and 6. Adding a few covariates has little effect on the estimates, as can be seen in Panel

Table 4.6.1: 2SLS, Abadie, and bivariate probit estimates of the effects of a third child on female labor supply

	2SLS	Abadie Estimates		Bivariate probit		
		Linear	Probit	MFX	ATE	TOT
	(1)	(2)	(3)	(4)	(5)	(6)
			A. No Covariates			
Employment	-0.138	-0.138	-0.137	-0.138	-0.139	-0.139
	(0.029)	(0.030)	(0.030)	(0.029)	(0.029)	(0.029)
			B. Some covariates (no age controls)			
Employment	-0.132	-0.132	-0.131	-0.135	-0.135	-0.135
	(0.029)	(0.029)	(0.028)	(0.028)	(0.028)	(0.028)
			C. Some covariates plus age at first birth			
Employment	-0.129	-0.129	-0.129	-0.133	-0.133	-0.133
	(0.028)	(0.028)	(0.028)	(0.026)	(0.026)	(0.026)
			D. Some covariates plus age at first birth and a dummy for age>30			
Employment	-0.124	-0.125	-0.125	-0.131	-0.131	-0.131
	(0.028)	(0.029)	(0.029)	(0.025)	(0.025)	(0.025)
			E. Some covariates plus age at first birth and age			
Employment	-0.120	-0.121	-0.121	-0.171	-0.171	-0.171
	(0.028)	(0.026)	(0.026)	(0.023)	(0.023)	(0.023)

Notes: Adapted from Angrist (2001). The table compares 2SLS estimates to alternative IV-type estimates of the effect of childbearing on labor supply using nonlinear models. Standard errors for the Abadie estimates were bootstrapped using 100 replications of subsamples of size 20,000. MFX denotes marginal effects; ATE is the average treatment effect; TOT is the average effect of treatment on the treated.

B. In this case, the covariates are all dummy variables, three for race (black, Hispanic, and other), and two indicating first and second-born boys (the excluded instrument is the interaction of these two). Panels C and D show that adding a linear term in age at first birth and a dummy for maternal age also leaves the estimates unchanged.

The invariance to covariates seems desirable: since the same-sex instrument is essentially independent of the covariates, control for covariates is unnecessary to eliminate bias and should primarily affect precision. Yet, as Panel E shows, the marginal effects generated by bivariate Probit are sensitive to the list of covariates. Swapping a dummy indicating mothers over 30 with a linear age term increases the bivariate Probit estimates markedly, to -.171, while leaving 2SLS and the Abadie estimators unchanged. This probably reflects the fact that the linear age change induces an extrapolation into cells where there is little data. Although there is no harm in reporting the results in Panel E, it's hard to see why the more robust 2SLS and Abadie estimators should not be featured as most likely more reliable.[40]

[40] Angrist (2001) makes the same point using twins instruments, and reports a similar pattern in a comparison of 2SLS,

4.6.4 The Bias of 2SLS★

It is a fortunate fact that the OLS estimator is not only consistent, it is also unbiased. This means that in a sample of any size, the estimated OLS coefficient vector has a distribution that is centered on the population coefficient vector.[41] The 2SLS estimator, in contrast, is consistent, but biased. This means that the 2SLS estimator only promises to be close the causal effect of interest in large samples. In small samples, the 2SLS estimator can differ systematically from the population estimand.

For many years, applied researchers have lived with the knowledge that 2SLS is biased without losing too much sleep. Neither of us heard much about the bias of 2SLS in our graduate econometrics classes. A series of papers in the early 1990s changed this, however. These papers show that 2SLS estimates can be highly misleading in cases relevant for empirical practice.[42]

The 2SLS estimator is most biased when the instruments are "weak," meaning the correlation with endogenous regressors is low, and when there are many over-identifying restrictions. When the instruments are both many and weak, the 2SLS estimator is biased towards the probability limit of the corresponding OLS estimate. In the worst-case scenario for many weak instruments, when the instruments are so weak that there really is no first-stage in the population, the 2SLS sampling distribution is centered on the probability limit of OLS. The theory behind this result is a little technical but the basic idea is easy to see. The source of the bias in 2SLS estimates is the randomness in estimates of the first-stage fitted values. In practice, the first-stage estimates reflect some of the randomness in the endogenous variable since the first-stage coefficients come from a regression of the endogenous variable on the instruments. If the population first-stage is zero, then all of the randomness in the first stage is due to the endogenous variable. This randomness turns into finite-sample correlation between first-stage fitted values and the second-stage errors, since the endogenous variable is correlated with the second-stage errors (or else you wouldn't be instrumenting in the first place).

A more formal derivation of 2SLS bias goes like this. To streamline the discussion we use matrices and vectors and a simple constant-effects model (it's difficult to discuss bias in a heterogeneous effects world, since the target parameter may be variable across estimators). Suppose you are interested in estimating the effect of a single endogenous regressor, stored in a vector x, on a dependent variable, stored in the vector y, with no other covariates. The causal model of interest can then be written

$$y = \beta x + \eta. \tag{4.6.17}$$

Abadie, and nonlinear structural estimates of models for hours worked. Angrist (1991) compares 2SLS and bivariate Probit estimates in sampling experiments.

[41] A more precise statement is that OLS is unbiased when, either (a) the CEF is linear or, (b) the regressors are non-stochastic, i.e., fixed in repeated samples. In practice, these qualifications do not seem to matter much. As a rule, the sampling distribution of $\hat{\beta} = \left[\sum_i \mathrm{X}_i\mathrm{X}_i'\right]^{-1}\sum_i \mathrm{X}_i\mathrm{Y}_i$, tends to be centered on the population analog, $\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$ in samples of any size, whether or not the CEF is linear or the regressors are stochastic.

[42] Key references are Nelson and Startz, (1990a,b); Buse (1992), Bekker (1994); and especially Bound, Jaeger, and Baker (1995).

The $N \times \mathrm{Q}$ matrix of instrumental variables is Z, with the associated first-stage equation

$$x = Z\pi + \xi. \tag{4.6.18}$$

OLS estimates of (4.6.17) are biased because η_i is correlated with ξ_i. The instruments, Z_i are uncorrelated with ξ_i by construction and uncorrelated with η_i by assumption.

The 2SLS estimator is

$$\widehat{\beta}_{2SLS} = (x'P_Z x)^{-1} x'P_Z y = \beta + (x'P_Z x)^{-1} x'P_Z \eta.$$

where $P_Z = Z(Z'Z)^{-1}Z'$ is the projection matrix that produces fitted values from a regression of x on Z. Substituting for x in $x'P_Z\eta$, we get

$$\widehat{\beta}_{2SLS} - \beta = (x'P_Z x)^{-1} \left(\pi'Z' + \xi'\right) P_Z \eta = (x'P_Z x)^{-1} \pi'Z'\eta + (x'P_Z x)^{-1} \xi'P_Z\eta \tag{4.6.19}$$

The bias in 2SLS comes from the nonzero expectation of terms on the right hand side.

The expectation of (4.6.19) is hard to evaluate because the expectation operator does not pass through the inverse $(x'P_Z x)^{-1}$, a nonlinear function. It's possible to show, however, that the expectation of the ratios on the right hand side of (4.6.19) can be closely approximated by the ratio of expectations. In other words,

$$E[\widehat{\beta}_{2SLS} - \beta] \approx (E[x'P_Z x])^{-1} E[\pi'Z'\eta] + (E[x'P_Z x])^{-1} E[\xi'P_Z\eta].$$

This approximation is much better than the usual first-order asymptotic approximation invoked in large-sample theory, so we think of it as giving us a good measure of the finite-sample behavior of the 2SLS estimator.[43] Furthermore, because $E[\pi'Z'\xi] = 0$ and $E[\pi'Z'\eta] = 0$, we have

$$E[\widehat{\beta}_{2SLS} - \beta] \approx \left[E\left(\pi'Z'Z\pi\right) + E(\xi'P_Z\xi)\right]^{-1} E\left(\xi'P_Z\eta\right). \tag{4.6.20}$$

The approximate bias of 2SLS therefore comes from the fact that $E\left(\xi'P_Z\eta\right)$ is not zero unless η_i and ξ_i are uncorrelated. But correlation between η_i and ξ_i is what led us to use IV in the first place.

Further manipulation of (4.6.20) generates an expression that is especially useful:

$$E[\widehat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\eta\xi}}{\sigma_\xi^2} \left[\frac{E\left(\pi'Z'Z\pi\right)/\mathrm{Q}}{\sigma_\xi^2} + 1\right]^{-1}$$

[43] See Bekker (1994) and Angrist and Krueger (1995). This is also called a group-asymptotic approximation because it can be derived from an an asymptotic sequence that lets the *number instruments* go to infinity at the same time as the number of observations goes to infinity, thereby keeping the number of observations per instrument constant.

(see the appendix for a derivation). The term $(1/\sigma_\xi^2)E\left(\pi' Z' Z \pi\right)/\mathrm{Q}$ is the F-statistic for the joint significance of all regressors in the first stage regression.[44] Call this statistic F, so that we can write

$$E[\widehat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\eta\xi}}{\sigma_\xi^2} \frac{1}{F+1}. \tag{4.6.21}$$

From this we see that as the first stage F-statistic gets small, the bias of 2SLS approaches $\frac{\sigma_{\eta\xi}}{\sigma_\xi^2}$. The bias of the OLS estimator is $\frac{\sigma_{\eta\xi}}{\sigma_x^2}$, which also equals $\frac{\sigma_{\eta\xi}}{\sigma_\xi^2}$ if $\pi = 0$. Thus, we have shown that 2SLS is centered on the same point as OLS when the first stage is zero. More generally, we can say 2SLS estimates are "biased towards" OLS estimates when there isn't much of a first stage. On the other hand, the bias of 2SLS vanishes when F gets large, as it should happen in large samples when $\pi \neq 0$.

When the instruments are weak, the F-statistic itself varies inversely with the number of instruments. To see why, consider adding useless instruments to your 2SLS model, that is, instruments with no effect on the first-stage R-squared. The model sum of squares, $E\left(\pi' Z' Z \pi\right)$, and the residual variance, σ_ξ^2, will both stay the same while Q goes up. The F-statistic becomes smaller as a result. From this we learn that the addition of many weak instruments increases bias.

Intuitively, the bias in 2SLS is a consequence of the fact that the first stage is estimated. If the first stage coefficients were known, we could use $\widehat{x}_{pop} = Z\pi$ for the first-stage fitted values. These fitted values are uncorrelated with the second stage error. In practice, however, we use $\widehat{x} = P_Z x = Z\pi + P_Z\xi$, which differs from $\widehat{x}_{pop}$ by the term $P_Z\xi$. The bias in 2SLS arises from the fact that $P_Z\xi$ is correlated with η, so some of the correlation between errors in the first and second stages seeps in to our 2SLS estimates through the sampling variability in $\widehat{\pi}$. Asymptotically, this correlation is negligible, but real life does not play out in "asymptopia".

The bias formula, (4.6.21), shows that the bias in 2SLS is an increasing function of the number of instruments, so clearly bias is least in the just-identified case when the number of instruments is as low as it can get. It turns out, however, that just-identified 2SLS (say, the simple Wald estimator) is approximately *unbiased*. This is hard to show formally because just-identified 2SLS has no moments (i.e., the sampling distribution has fat tails). Nevertheless, even with weak instruments, just-identified 2SLS is approximately centered where it should be (we therefore say that just-identified 2SLS is median-unbiased). This is not to say that you can happily use weak instruments in just-identified models. With a weak instrument, just-identified IV estimates tend to be highly unstable and imprecise.

The LIML estimator is approximately median-unbiased for over-identified constant-effects models, and therefore provides an attractive alternative to just-identified estimation using one instrument at a time (see, e.g., Davidson and MacKinnon, 1993, and Mariano, 2001). LIML has the advantage of having the same

[44] Sort of; the actual F-statistic is $(1/\hat{\sigma}_\xi^2)\hat{\pi}' Z' Z \hat{\pi}/\mathrm{Q}$, where hats denote estimates. $(1/\sigma_\xi^2)E\left(\pi' Z' Z \pi\right)/\mathrm{Q}$ is therefore sometimes called the population F-statistic since it's the F-statistic we'd get in an infinitely large sample. In practice, the distinction between population and sample F matters little in this context.

large-sample distribution as 2SLS (under constant effects) while providing finite-sample bias reduction. A number of estimators reduce the bias in overidentified 2SLS models. But an extensive Monte Carlo study by Flores-Lagunes (2007) suggests that LIML does at least as well as the alternatives in a wide range of circumstances (in terms of bias, mean absolute error, and the empirical rejection rates for t-tests). Another advantage of LIML is that many statistical packages compute it while other estimators typically require some programming.[45]

We use a small Monte Carlo experiment to illustrate some of the theoretical results from the discussion above. The simulated data are drawn from the following model,

$$\begin{aligned} y_i &= \beta x_i + \eta_i \\ x_i &= \sum_{j=1}^{\mathrm{Q}} \pi_j z_{ij} + \xi_i \end{aligned}$$

with $\beta = 1$, $\pi_1 = 0.1$, $\pi_j = 0 \; \forall j > 1$,

$$\left(\begin{array}{c} \eta_i \\ \xi_i \end{array} \right) \Bigg| Z \sim N \left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} 1 & 0.8 \\ 0.8 & 1 \end{array} \right) \right),$$

where the z_{ij} are independent, normally distributed random variables with mean zero and unit variance. The sample size is 1000.

Figure 4.6.1 shows the Monte Carlo distributions of four estimators: OLS, just identified IV (i.e. 2SLS with Q= 1, labeled IV), 2SLS with two instruments (for Q= 2, labeled 2SLS), and LIML with Q= 2. The OLS estimator is biased and centered around a value of about 1.79. IV is centered around 1, the value of β. 2SLS with one weak and one uninformative instrument is moderately biased towards OLS (the median is 1.07). The distribution function for LIML with Q= 2 is basically indistinguishable from that for just-identified IV, even though the LIML estimator uses a completely uninformative instrument.

Figure 4.6.2 reports simulation results where we set Q= 20. Thus, in addition to the one informative but weak instrument, we added 19 worthless instruments. The figure again shows OLS, 2SLS, and LIML distributions. The bias in 2SLS is now much worse (the median is 1.53, close to the OLS median). The sampling distribution of the 2SLS estimator is also much tighter than in the Q= 2 case. LIML continues to

[45] LIML is available in SAS and in STATA 10. With weak instruments, LIML standard errors are not quite right, but Bekker (1994) gives a simple fix for this. Why is LIML unbiased? Expression (4.6.21) shows that the approximate bias of 2SLS is proportional to the bias of OLS. From this we conclude that there is a linear combination of OLS and 2SLS that is approximately unbiased. LIML turns out to be just such a "combination estimator". Like the bias of 2SLS, the approximate unbiasedness of LIML can be shown using a Bekker-style group-asymptotic sequence that fixes the ratio of instruments to sample size. Its worth mentioning, however, that LIML is biased in models with a certain type of heteroskedasticity; See Hausman, Newey, and Wouterson (2006) for details.

perform well and is centered around $\beta = 1$, with a bit more dispersion than in the Q= 2 case.

Finally, Figure 4.6.3 reports simulation results from a model that is truly unidentified. In this case, we set $\pi_j = 0$; $j = 1, ..., 20$. Not surprisingly, all the sampling distributions are centered around the same value as OLS. On the other hand, the 2SLS sampling distribution is much tighter than the LIML distribution. We would say advantage-LIML in this case because the widely dispersed LIML sampling distribution correctly reflects the fact that the sample is uninformative about the parameter of interest.

What does this mean in practice? Besides retaining a vague sense of worry about your first stage, we recommend the following:

1. Report the first stage and think about whether it makes sense. Are the magnitude and sign as you would expect, or are the estimates too big or large but wrong-signed? If so, perhaps your hypothesized first-stage mechanism isn't really there, rather, you simply got lucky.

2. Report the F-statistic on the excluded instruments. The bigger this is, the better. Stock, Wright, and Yogo (2002) suggest that F-statistics above about 10 put you in the safe zone though obviously this cannot be a theorem.

3. Pick your best single instrument and report just-identified estimates using this one only. Just-identified IV is median-unbiased and therefore unlikely to be subject to a weak-instruments critique.

4. Check over-identified 2SLS estimates with LIML. LIML is less precise than 2SLS but also less biased. If the results come out similar, be happy. If not, worry, and try to find stronger instruments.

5. Look at the coefficients, t-statistics, and F-statistics for excluded instruments in the reduced-form regression of dependent variables on instruments. Remember that the reduced form is proportional to the causal effect of interest. Most importantly, the reduced-form estimates, since they are OLS, are unbiased. As Angrist and Krueger (2001) note, if you can't see the causal relation of interest in the reduced form, it's probably not there.[46]

We illustrate some of this reasoning in a re-analysis of the Angrist and Krueger (1991) quarter-of-birth study. Bound, Jaeger, and Baker (1995) argued that bias is a major concern when using quarter birth as an instrument for schooling, in spite of the fact that sample size exceeds 300,000. "Small sample" is clearly relative. Earlier in the chapter, we saw that the QOB pattern in schooling is clearly reflected in the reduced form, so there would seem to be little cause for concern. On the other hand, Bound, Jaeger, and Baker (1995) argue that the most relevant models have additional controls not included in these reduced forms. Table 4.6.2 reproduces some of the specifications from Angrist and Krueger (1991) as well as other specifications in the spirit of Bound, Jaeger, and Baker (1995).

[46] A recent paper by Chernozhukov and Hansen (2007) formalizes this maxim.

Table 4.6.2: Alternative IV estimates of the economic returns to schooling

	(1)	(2)	(3)	(4)	(5)	(6)
2SLS	0.105	0.435	0.089	0.076	0.093	0.091
	(0.020)	(0.450)	(0.016)	(0.029)	(0.009)	(0.011)
LIML	0.106	0.539	0.093	0.081	0.106	0.110
	(0.020)	(0.627)	(0.018)	(0.041)	(0.012)	(0.015)
F-statistic (excluded instruments)	32.27	0.42	4.91	1.61	2.58	1.97
	Controls					
Year of birth	✓	✓	✓	✓	✓	✓
State of birth					✓	✓
Age, Age squared		✓		✓		✓
	Excluded Instruments					
Quarter of birth	✓	✓				
Quarter of birth*year of birth			✓	✓	✓	✓
Quarter of birth*state of birth					✓	✓
Number of excluded instruments	3	2	30	28	180	178

Notes: The table compares 2SLS and LIML estimates using alternative sets of instruments and controls. The OLS estimate corresponding to the models reported in columns 1-4 is .071; the OLS estimate corresponding to the models reported in columns 5-6 is .067. Data are from the Angrist and Krueger (1991) 1980 Census sample. The sample size is 329,509. Standard errors are reported in parentheses.

The first column in the table reports 2SLS and LIML estimates of a model using three quarter of birth dummies as instruments with year of birth dummies as covariates. The OLS estimate for this specification is 0.071, while the 2SLS estimate is a bit higher at 0.105. The first-stage F-statistic is over 32, well above the danger zone. Not surprisingly, the LIML estimate is almost identical to 2SLS in this case.

Angrist and Krueger (1991) experimented with models that include age and age squared measured in quarters as additional controls. These controls are meant to pick up omitted age effects that might confound the quarter-of-birth instruments. The addition of age and age squared reduces the number of instruments to two, since age in quarters, year of birth, and quarter of birth are linearly dependent. As shown in column 2, the first stage F-statistic drops to 0.4 when age and age squared are included as controls, a sure sign of trouble. But the 2SLS standard error is high enough that we would not draw any substantive conclusions from this estimate. The LIML estimate is even less precise. This model is effectively unidentified.

Columns 3 and 4 report the results of adding interactions between quarter of birth dummies and year of birth dummies to the instrument list, so that there are 30 instruments, or 28 when the age and age squared variables are included. The first stage F-statistics are 4.9 and 1.6 in these two specifications. The 2SLS estimates are a bit lower than in column 1 and hence closer to OLS. But LIML is not too far away from 2SLS. Although the LIML standard error is pretty big in column 4, it is not so large that the estimate is uninformative. On balance, there seems to be little cause for worry about weak instruments, even with the age quadratic included.

The most worrisome specifications are those reported in columns 5 and 6. These estimates were produced by adding 150 interactions between quarter of birth and state of birth to the 30 interactions between quarter of birth and year of birth. The rationale for the inclusion of state-of-birth interactions in the instrument list is to exploit differences in compulsory schooling laws across states. But this leads to highly over-identified models with 180 (or 178) instruments, many of which are weak. The first stage F-statistics for these models are 2.6 and 2.0, well into the discomfort zone. On the plus side, the LIML estimates again look fairly similar to 2SLS. Moreover, the LIML standard errors differ little from the 2SLS standard errors in this case. This suggests that you can't always determine instrument relevance using a mechanical rule such as "F>10". In some cases, a low F may not be fatal.[47]

Finally, it's worth noting that in applications with multiple endogenous variables, the conventional first-stage F is no longer appropriate. To see why, suppose there are two instruments for two endogenous variables and that the first instrument is strong and predicts both endogenous variables well while the second instrument is weak. The first-stage F-statistics in each of the two first stage equations are likely to be high but the model is weakly identified because one instrument is not enough to capture two causal effects. A simple modification of the first-stage F for this case is given in the appendix.

[47] Cruz and Moreira (2005) similarly conclude that, low F-statistics notwithstanding, there is little bias in the Angrist and Krueger (1991) 180-instrument specifications.

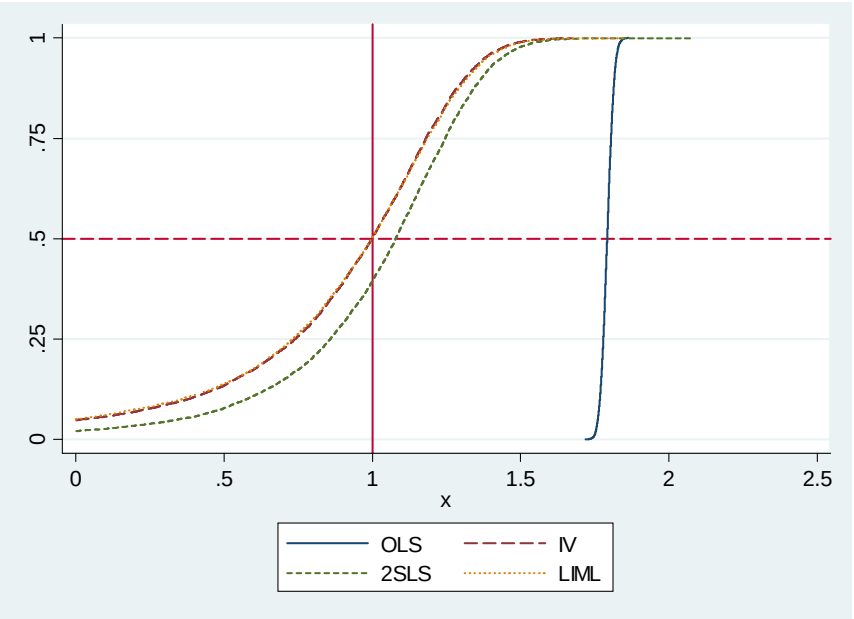

Figure 4.6.1: Distribution of the OLS, IV, 2SLS, and LIML estimators. IV uses one instrument, while 2SLS and LIML use two instruments.

4.7 Appendix

Derivation of Equation (4.6.8)

Rewrite equation (4.6.7) as follows

$$Y_{ij} = \mu^* + \pi_0 \tau_i + (\pi_0 + \pi_1)\overline{S}_j + \nu_i;$$

where $\tau_i \equiv \mathrm{s}_i - \overline{S}_j$. Since τ_i and $\overline{S}_j$ are uncorrelated by construction, we have:

$$\begin{aligned} \rho_1 &= \pi_0 + \pi_1. \\ \pi_0 &= \frac{C(\tau_i, Y_{ij})}{V(\tau_i)}. \end{aligned}$$

Simplifying the second line,

$$\begin{aligned} \pi_0 &= \frac{C[(s_i - \overline{S}_j), Y_{ij}]}{[V(\mathrm{s}_i) - V(\overline{S}_j)]} \\ &= \left[\frac{C(\mathrm{s}_i, Y_{ij})}{V(\mathrm{s}_i)}\right]\left[\frac{V(\mathrm{s}_i)}{V(\mathrm{s}_i) - V(\overline{S}_j)}\right] - \left[\frac{C(\overline{S}_j, Y_{ij})}{V(\overline{S}_j)}\right]\left[\frac{V(\overline{S}_j)}{V(\mathrm{s}_i) - V(\overline{S}_j)}\right] \\ &= \rho_0\phi + \rho_1(1-\phi) = \rho_1 + \phi(\rho_0 - \rho_1) \end{aligned}$$

where $\phi \equiv \frac{V(s_i)}{V(\mathrm{s}_i) - V(\overline{S}_j)}$. Solving for π_1, we have

$$\pi_1 = \rho_1 - \pi_0 = \phi(\rho_1 - \rho_0).$$

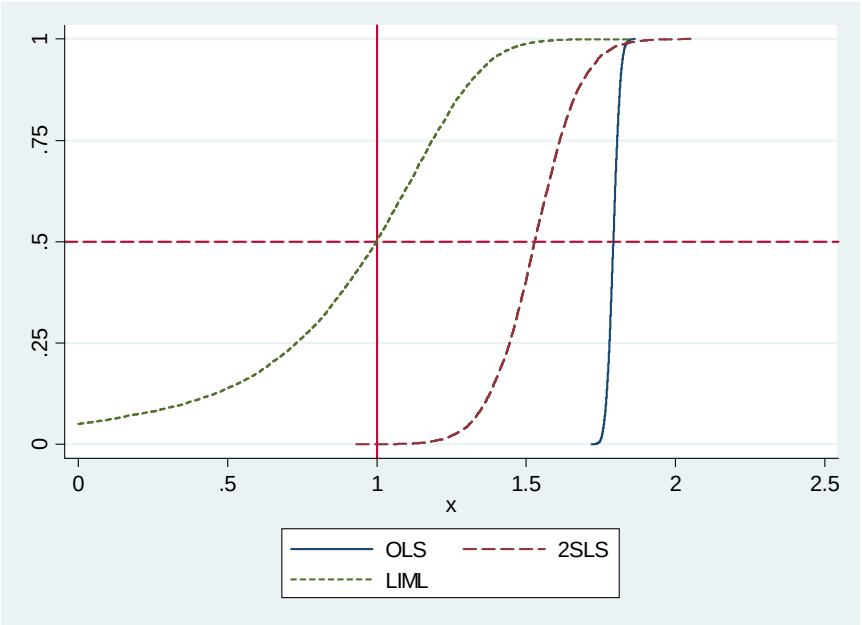

Figure 4.6.2: Distribution of the OLS, 2SLS, and LIML estimators with 20 instruments

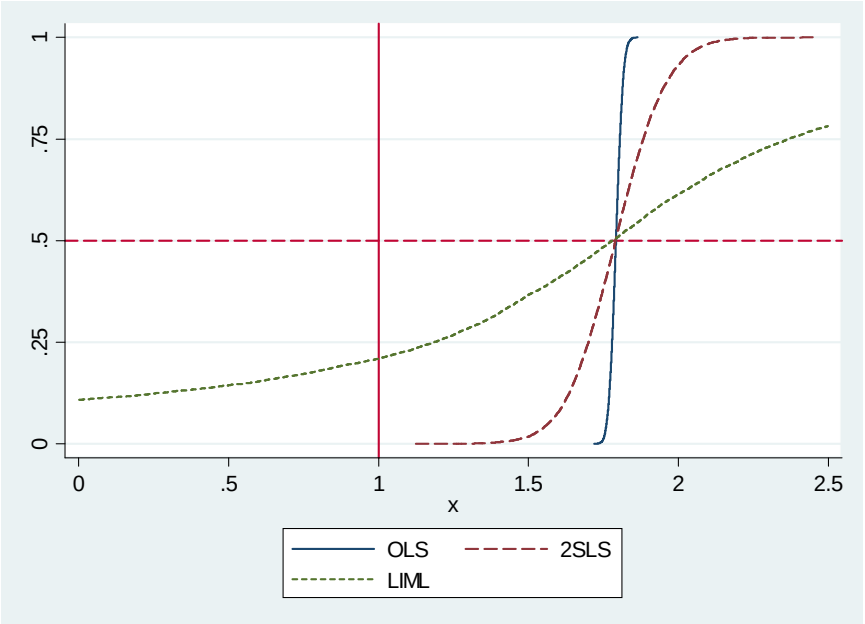

Figure 4.6.3: Distribution of the OLS, 2SLS, and LIML estimators with 20 worthless instruments

Derivation of the approximate bias of 2SLS

Start from the last equality in (4.6.20):

$$E[\widehat{\beta}_{2SLS} - \beta] \approx \left[E\left(\pi' Z' Z \pi\right) + E\left(\xi' P_Z \xi\right)\right]^{-1} E\left(\xi' P_Z \eta\right).$$

The magic of linear algebra helps us simplify this expression: The term $\xi' P_Z \xi$ is a scalar and therefore equal to its trace; the trace is a linear operator which passes through expectations and is invariant to cyclic permutations; finally, the trace of P_Z, an idempotent matrix, is equal to it's rank, Q. Using these facts, we have

$$\begin{aligned} E\left(\xi' P_Z \xi\right) &= E\left[\mathrm{tr}\left(\xi' P_Z \xi\right)\right] \\ &= E\left[\mathrm{tr}\left(P_Z \xi \xi'\right)\right] \\ &= \mathrm{tr}\left(P_Z E\left[\xi \xi'\right]\right) \\ &= \mathrm{tr}\left(P_Z \sigma_\xi^2 I\right) \\ &= \sigma_\xi^2 \mathrm{tr}\left(P_Z\right) \\ &= \sigma_\xi^2 \mathrm{Q}, \end{aligned}$$

where we have assumed that ξ_i is homoskedastic. Similarly, applying the trace trick to $\xi' P_Z \eta$ shows that this term is equal to $\sigma_{\eta\xi}\mathrm{Q}$. Therefore,

$$\begin{aligned} E[\widehat{\beta}_{2SLS} - \beta] &\approx \left[E\left(\pi' Z' Z \pi\right) + \sigma_\xi^2 \mathrm{Q}\right]^{-1} E\left[\mathrm{tr}\left(\xi' P_Z \eta\right)\right] \\ &= \sigma_{\eta\xi}\mathrm{Q}\left[E\left(\pi' Z' Z \pi\right) + \sigma_\xi^2 \mathrm{Q}\right]^{-1} \\ &= \frac{\sigma_{\eta\xi}}{\sigma_\xi^2}\left[\frac{E\left(\pi' Z' Z \pi\right)/\mathrm{Q}}{\sigma_\xi^2} + 1\right]^{-1}. \end{aligned}$$

Multivariate first-stage F-statistics

Assume any exogenous covariates have been partialled out of the instrument list and that there are two endogenous variables, x_1 and x_2 with coefficients δ_1 and δ_2. We are interested in the bias of the 2SLS estimator of δ_2 when x_1 is also treated as endogenous. The second stage equation is

$$y = P_Z x_1 \delta_1 + P_Z x_2 \delta_2 + [\eta + (x_1 - P_Z x_1)\delta_1 + (x_2 - P_Z x_2)\delta_2]. \tag{4.7.1}$$

where $P_Z x_1$ and $P_Z x_2$ are the first-stage fitted values from regressions of x_1 and x_2 on Z. By the usual anatomy formula for multivariate regression, δ_2 in (4.7.1) is the bivariate regression of y on the residual from

a regression of $P_Z x_2$ on $P_Z x_1$. This residual is

$$[I - P_Z x_1 (x_1' P_Z x_1)^{-1} x_1' P_Z] P_Z x_2 = M_{1z} P_Z x_2,$$

where $M_{1z} = [I - P_Z x_1 (x_1' P_Z x_1)^{-1} x_1' P_Z]$ is the relevant residual-maker matrix. In addition, note that $M_{1z} P_Z x_2 = P_Z [M_{1z} x_2]$.

From here we conclude that the 2SLS estimator of δ_2 is the OLS regression on $P_Z[M_{1z} x_2]$, in other words, OLS on the fitted values from a regression of $M_{1z} x_2$ on Z. This is the same as 2SLS using P_Z to instrument $M_{1z} x_2$. So the 2SLS estimator of δ_2 can be written

$$[x_2' M_{1z} P_Z M_{1z} x_2]^{-1} x_2' M_{1z} P_Z y = \delta_2 + [x_2' M_{1z} P_Z M_{1z} x_2]^{-1} x_2' M_{1z} P_Z \eta.$$

The explained sum of squares (numerator of the F-statistic) that determines the bias of the 2SLS estimator of δ_2 is therefore the expectation of $[x_2' M_{1z} P_Z M_{1z} x_2]$, while the bias comes from the fact that the expectation $E[\xi' M_{1z} P_Z \eta]$ is non-zero when η and ξ are correlated.

Here's how to compute this F-statistic in practice: (a) Regress the first stage fitted values for the regressor of interest, $P_Z x_2$, on the other first-stage fitted values and any exogenous covariates. Save the residuals from this step; (b) Construct the F-statistic for excluded instruments in a first-stage regression of the residuals from (a) on the excluded instruments. Note that you should get the 2SLS coefficient of interest in a 2SLS procedure where the residuals from (a) are instrumented using Z, with no other covariates or endogenous variables. Use this fact to check your calculation.

Chapter 5

Parallel Worlds: Fixed Effects, Differences-in-differences, and Panel Data

> The first thing to realize about parallel universes . . . is that they are not parallel.
>
> Douglas Adams, *Mostly Harmless* (1995)

The key to causal inference in chapter 3 is control for observed confounding factors. If important confounders are unobserved, we might try to get at causal effects using IV as discussed in Chapter 4. Good instruments are hard to find, however, so we'd like to have other tools to deal with unobserved confounders. This chapter considers a variation on the control theme: strategies that use data with a time or cohort dimension to control for unobserved-but-fixed omitted variables. These strategies punt on comparisons in levels, while requiring the counterfactual *trend* behavior of treatment and control groups to be the same. We also discuss the idea of controlling for lagged dependent variables, another strategy that exploits timing.

5.1 Individual Fixed Effects

One of the oldest questions in Labor Economics is the connection between union membership and wages. Do workers whose wages are set by collective bargaining earn more because of this, or would they earn more anyway? (Perhaps because they are more experienced or skilled). To set this question up, let Y_{it} equal the (log) earnings of worker i at time t and let D_{it} denote his union status. The observed Y_{it} is either Y_{0it} or Y_{1it}, depending on union status. Suppose further that

$$E(\text{Y}_{0it}|A_i, \text{X}_{it}, t, \text{D}_{it}) = E(\text{Y}_{0it}|A_i, \text{X}_{it}, t),$$

i.e. union status is as good as randomly assigned conditional on unobserved worker ability, A_i, and other observed covariates X_{it}, like age and schooling.

The key to fixed-effects estimation is the assumption that the unobserved A_i appears without a time subscript in a linear model for $E(\mathrm{Y}_{0it}|A_i,\mathrm{X}_{it},t)$:

$$E(\mathrm{Y}_{0it}|A_i,\mathrm{X}_{it},t) = \alpha + \lambda_t + A_i'\gamma + \mathrm{X}_{it}\delta, \tag{5.1.1}$$

Finally, we assume that the causal effect of union membership is additive and constant:

$$E(\mathrm{Y}_{1it}|A_i,\mathrm{X}_{it},t) = E(\mathrm{Y}_{0it}|A_i,\mathrm{X}_{it},t) + \rho.$$

This implies

$$E(\mathrm{Y}_{it}|A_i,\mathrm{X}_{it},t,\mathrm{D}_{it}) = \alpha + \lambda_t + \rho\mathrm{D}_{it} + A_i'\gamma + \mathrm{X}_{it}\delta, \tag{5.1.2}$$

where ρ is the causal effect of interest. The set of assumptions leading to (5.1.2) is more restrictive those we used to motivate regression in Chapter 3; we need the linear, additive functional form to make headway on the problem of *unobserved* confounders using panel data with no instruments.[1]

Equation (5.1.2) implies

$$\mathrm{Y}_{it} = \alpha_i + \lambda_t + \rho\mathrm{D}_{it} + \mathrm{X}_{it}\delta + \varepsilon_{it}. \tag{5.1.3}$$

where

$$\alpha_i \equiv \alpha + A_i'\gamma.$$

This is a *fixed-effects model.* Given panel data, i.e., repeated observations on individuals, the causal effect of union status on wages can be estimated by treating α_i, the fixed effect, as a parameter to be estimated. The *year effect*, λ_t, is also treated as a parameter to be estimated. The unobserved individual effects are coefficients on dummies for each individual while the year effects are coefficients on time dummies.[2]

It might seem like there are an awful lot of parameters to be estimated in the fixed effects model. For

[1] In some cases, we can allow heterogeneous treatment effects so that

$$E(\mathrm{Y}_{1it} - \mathrm{Y}_{0it}|A_i,\mathrm{X}_{it},t) = \rho_i.$$

See, e.g., Wooldridge (2005), who discusses estimators for the average of ρ_i.

[2] An alternative to the fixed-effects specification is "random effects" (See, e.g., Wooldridge, 2006). The random-effects model assumes that α_i is uncorrelated with the regressors. Because the omitted variable in a random-effects model is uncorrelated with included regressors there is no bias from ignoring it - in effect, it becomes part of the residual. The most important consequence of random effects is that the residuals for a given person are correlated across periods. Chapter 8 discusses the implications of this for standard errors. An alternative approach is GLS, which promises to be more efficient if the assumptions of the random-effects model are satisfied (linear CEF, homoskedasticity). We prefer OLS/fix-the-standard-errors to GLS under random-effects assumptions. As discussed in Section 3.4.1, GLS requires stronger assumptions than those we are comfortable with and the resulting efficiency gain is likely to be modest.

example, the Panel Survey of Income Dynamics, a widely-used panel data set, includes about 5,000 working-age men observed for about 20 years. So there are roughly 5,000 fixed effects. In practice, however, this doesn't matter. Treating the individual effects as parameters to be estimated is algebraically the same as estimation in deviations from means. In other words, first we calculate the individual averages

$$\overline{\mathrm{Y}}_i = \alpha_i + \overline{\lambda} + \rho\overline{\mathrm{D}}_i + \overline{\mathrm{X}}_i\delta + \overline{\varepsilon}_i.$$

Subtracting this from (5.1.3) gives

$$\mathrm{Y}_{it} - \overline{\mathrm{Y}}_i = \lambda_t - \overline{\lambda} + \rho\left(\mathrm{D}_{it} - \overline{\mathrm{D}}_i\right) + \left(\mathrm{X}_{it} - \overline{\mathrm{X}}_i\right)\delta + \left(\varepsilon_{it} - \overline{\varepsilon}_i\right), \tag{5.1.4}$$

so deviations from means kills the unobserved individual effects.[3]

An alternative to deviations from means is differencing. In other words, we estimate,

$$\Delta\mathrm{Y}_{it} = \Delta\lambda_t + \rho\Delta\mathrm{D}_{it} + \Delta\mathrm{X}_{it}\delta + \Delta\varepsilon_{it}, \tag{5.1.5}$$

where the Δ prefix denotes the change from one year to the next. For example, $\Delta\mathrm{Y}_{it} = \mathrm{Y}_{it} - \mathrm{Y}_{it-1}$. With two periods, differencing is algebraically the same as deviations from means, but not otherwise. Both should work, although with homoskedastic and serially uncorrelated ε_{it} deviations from means is more efficient. You might find differencing more convenient if you have to do it by hand, though the differenced standard errors should be adjusted for the fact that the differenced residuals are serially correlated.

Some regression packages automate the deviations-from-means estimator, with an appropriate standard-error adjustment for the degrees of freedoms lost in estimating N individual means. This is all that's needed to get the standard errors right with a homoskedastic, serially uncorrelated residual. The deviations-from-means estimator has many names, including the "within estimator" and "analysis of covariance". Estimation in deviations-from-means form is also called *absorbing* the fixed effects.[4]

Freeman (1984) uses four data sets to estimate union wage effects under the assumption that selection into union status is based on unobserved-but-fixed individual characteristics. Table 5.1.1 displays some of his estimates. For each data set, the table displays results from a fixed-effects estimator and the corresponding cross-section estimates. The cross section estimates are typically higher (ranging from .15-.25) than the

[3] Why is deviations from means the same as estimating each fixed effect in (5.1.3)? Because, by the regression anatomy formula, (3.1.3), any set of multivariate regression coefficients can be estimated in two steps. To get the multivariate coefficient on one set of variables, first regress them on all the other included variables, then regress the original dependent variable on the residuals from this first step. The residuals from a regression on a full set of person-dummies in a person-year panel are deviations from person means.

[4] The fixed effects are not estimated consistently in a panel where the number of periods T is fixed while $N \to \infty$. This is called the "incidental parameters problem," a name which reflects the fact that the number of parameters grows with the sample size. Nevertheless, other parameters in the fixed effects model - the ones we care about - are consistently estimated.

fixed effects estimates (ranging from .10-.20). This may indicate positive selection bias in the cross-section estimates, though selection bias is not the only explanation for the lower fixed-effects estimates.

Table 5.1.1: Estimated effects of union status on log wages

Survey	Cross section estimate	Fixed effects estimate
May CPS, 1974-75	0.19	0.09
National Longitudinal Survey of Young Men, 1970-78	0.28	0.19
Michigan PSID, 1970-79	0.23	0.14
QES, 1973-77	0.14	0.16

Notes: Adapted from Freeman (1984). The table reports cross-section and panel estimates of the union relative wage effect. The estimates were calculated using the surveys listed at left. The cross-section estimates include controls for demographic and human capital variables.

Although they control for a certain type of omitted variable, fixed-effects estimates are notoriously susceptible to attenuation bias from measurement error. On one hand, economic variables like union status tend to be persistent (a worker who is a union member this year is most likely a union member next year). On the other hand, measurement error often changes from year-to-year (union status may be misreported or miscoded this year but not next year). Therefore, while union status may be misreported or miscoded for only a few workers in any single year, the observed year-to-year changes in union status may be mostly noise. In other words, there is more measurement error in the regressors in an equation like (5.1.5) or (5.1.4) than in the levels of the regressors. This fact may account for smaller fixed-effects estimates.[5]

A variant on the measurement-error problem arises from that fact that the differencing and deviations-from-means estimators used to control for fixed effects typically remove both good and bad variation. In other words, these transformations may kill some of the omitted-variables-bias bathwater, but they also remove much of the useful information in the baby - the variable of interest. An example is the use of twins to estimate the causal effect of schooling on wages. Although there is no time dimension to this problem, the basic idea is the same as the union problem discussed above: twins have similar but largely unobserved family and genetic backgrounds. We can therefore control for their common family background by including a family fixed effect in samples of pairs of twins.

Ashenfelter and Krueger (1994) and Ashenfelter and Rouse (1998) estimate the returns to schooling using samples of twins, controlling for family fixed effects. Because there are two twins from each family, this is the same as regressing differences in earnings within twin-pairs on differences in their schooling. Surprisingly, the with-family estimates come our larger than OLS. But how do differences in schooling come about between individuals who are otherwise so much alike? Bound and Solon (1999) point out that there are small differences between twins, with first-borns typically having higher birth weight and higher IQ scores (here differences in birth timing are measured in minutes). While these within-twin differences are not large,

[5] See Griliches and Hausman (1986) for a more complete analysis of measurement error in panel data.

neither is the difference in their schooling. Hence, a small amount of unobserved ability differences among twins could be responsible for substantial bias in the resulting estimates.

What should be done about measurement error and related problems in models with fixed effects? A possible fix-up for measurement error is instrumental variables. Ashenfelter and Krueger (1994) use cross-sibling reports to construct instruments for schooling differences across twins. For example, they use each twin's report of his brother's schooling as an instrument for self-reports. A second approach is to bring in external information on the extent of measurement error and adjust naive estimates accordingly. In a study of union wage effects, Card (1996) uses external information from a separate validation survey to adjust panel-data estimates for measurement error in reported union status. But data from multiple reports and repeated measures of the sort used by Ashenfelter and Rouse (1994) and Card (1996) are unusual. At a minimum, therefore, it's important to avoid overly strong claims when interpreting fixed-effects estimates (never bad advice for an applied econometrician in any case).

5.2 Differences-in-differences: Pre and Post, Treatment and Control

The fixed effects strategy requires panel data, that is, repeated observations on the same individuals (or firms or whatever the unit of observation might be). Often, however, the regressor of interest varies only at a more aggregate level such as state or cohort. For example, state policies regarding health care benefits for pregnant workers or minimum wages change across states but not within states. The source of omitted variables bias when evaluating these policies must therefore be unobserved variables at the state and year level.

To make this concrete, suppose we are interested in the effect of the minimum wage on employment, a classic question in Labor Economics. In a competitive labor market, increases in the minimum wage move us up a downward-sloping demand curve. Higher minimums therefore reduce employment, perhaps hurting the very workers minimum-wage policies were designed to help. Card and Krueger (1994) use a dramatic change in the New Jersey state minimum wage to see if this is true.

On April 1, 1992, New Jersey raised the state minimum from $4.25 to $5.05. Card and Krueger collected data on employment at fast food restaurants in New Jersey in February 1992 and again in November 1992. These restaurants (Burger King, Wendy's, and so on) are big minimum-wage employers. Card and Krueger collected data from the same type of restaurants in eastern Pennsylvania, just across the Delaware river. The minimum wage in Pennsylvania stayed at $4.25 throughout this period. They used their data set to compute differences-in-differences (DD) estimates of the effects of the New Jersey minimum wage increase. That is, they compared the change in employment in New Jersey to the change in employment in Pennsylvania around the time New Jersey raised its minimum.

DD is a version of fixed-effects estimation using aggregate data.[6] To see this, let

$$\begin{aligned} \mathrm{Y}_{1ist} &= \text{fast food employment at restaurant } i \text{ and period } t \\ &\quad \text{if there is a high state minimum wage} \\ \mathrm{Y}_{0ist} &= \text{fast food employment at restaurant } i \text{ and period } t \\ &\quad \text{if there is a low state minimum wage} \end{aligned}$$

These are potential outcomes - in practice, we only get to see one or the other. Fort example, we see Y_{1ist} in New Jersey in November of 1992. The heart of the DD setup is an additive structure for potential outcomes in the no-treatment state. Specifically, we assume that

$$E(\mathrm{Y}_{0ist}|s,t) = \gamma_s + \lambda_t \tag{5.2.1}$$

where s denotes state (New Jersey or Pennsylvania) and t denotes period (February, before the minimum wage increase or November, after the increase). This equations says that in the absence of a minimum wage change, employment is determined by the sum of a time-invariant state effect and a year effect that is common across states. The additive state effect plays the role of the unobserved individual effect in the previous subsection.

Let D_{st} be a dummy for high-minimum-wage states, where states are index by s and observed in period t. Assuming that $E(\mathrm{Y}_{1ist} - \mathrm{Y}_{0ist}|s,t)$ is a constant, denoted β, we have:

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \beta \mathrm{D}_{st} + \varepsilon_{ist} \tag{5.2.2}$$

where $E(\varepsilon_{ist}|s,t) = 0$. From here, we get

$$E[\mathrm{Y}_{ist}|s = PA, t = Nov] - E(\mathrm{Y}_{ist}|s = PA, t = Feb) = \lambda_{Nov} - \lambda_{Feb}$$

and

$$E(\mathrm{Y}_{ist}|s = NJ, t = Nov) - E(\mathrm{Y}_{ist}|s = NJ, t = Feb) = \lambda_{Nov} - \lambda_{Feb} + \beta.$$

The population difference-in-differences,

$$[E(\mathrm{Y}_{ist}|s = PA, t = Nov) - E(\mathrm{Y}_{ist}|s = PA, t = Feb)]$$

$$- [E(\mathrm{Y}_{ist}|s = NJ, t = Nov) - E(\mathrm{Y}_{ist}|s = NJ, t = Feb)] = \beta,$$

[6] The DD idea is at least as old as IV. Kennan (1995) references a 1915 BLS report using DD to study the employment effects of the minimum wage (Obenauer and von der Nienburg, 1915).

is the causal effect of interest. This is easily estimated using the sample analog of the population means.

Table 5.2.1: Average employment per store before and after the New Jersey minimum wage increase

Variable		PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1.	FTE employment before,	23.33	20.44	-2.89
	all available observations	(1.35)	(0.51)	(1.44)
2.	FTE employment after,	21.17	21.03	-0.14
	all available observations	(0.94)	(0.52)	(1.07)
3.	Change in mean FTE	-2.16	0.59	2.76
	employment	(1.25)	(0.54)	(1.36)

Notes: Adapted from Card and Krueger (1994), Table 3. The table reports average full-time equivalent (FTE) employment at restaurants in Pennsylvania and New Jersey before and after a minimum wage increase in New Jersey. The sample consists of all stores with data on employment. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing. Standard errors are reported in parentheses

Table 5.2.1 (based on Table 3 in Card and Krueger, 1994) shows average employment at fast food restaurants in New Jersey and Pennsylvania before and after the change in the New Jersey minimum wage. There are four cells in the first two rows and columns, while the margins show state differences in each period, the changes over time in each state, and the difference-in-differences. Employment in Pennsylvania restaurants is somewhat higher than in New Jersey in February but falls by November. Employment in New Jersey, in contrast, increases slightly. These two changes produce a positive difference-in-differences, the opposite of what we might expect if a higher minimum wage pushes businesses up the labor demand curve.

How convincing is this evidence against the standard labor-demand story? The key identifying assumption here is that employment *trends* would be the same in both states in the absence of treatment. Treatment induces a deviation from this common trend, as illustrated in figure 5.2.1. Although the treatment and control states can differ, this difference in captured by the state fixed effect, which plays the same role as the unobserved individual effect in (5.1.3).[7]

The common trends assumption can be investigated using data on multiple periods. In an update of their

[7]The common trends assumption can be applied to transformed data, for example,

$$E(\log Y_{0ist}|s,t) = \gamma_s + \lambda_t.$$

Note, however, that if there is a common trend in logs, there will not be one in levels and vice versa. Athey and Imbens (2006) introduce a semi-parametric DD estimator that allows for common trends after an unknown transformation, which they propose to use the data to estimate. Poterba, Venti and Wise (1995) and Meyer, Viscusi, and Durbin (1995) discuss DD-type models for quantiles.

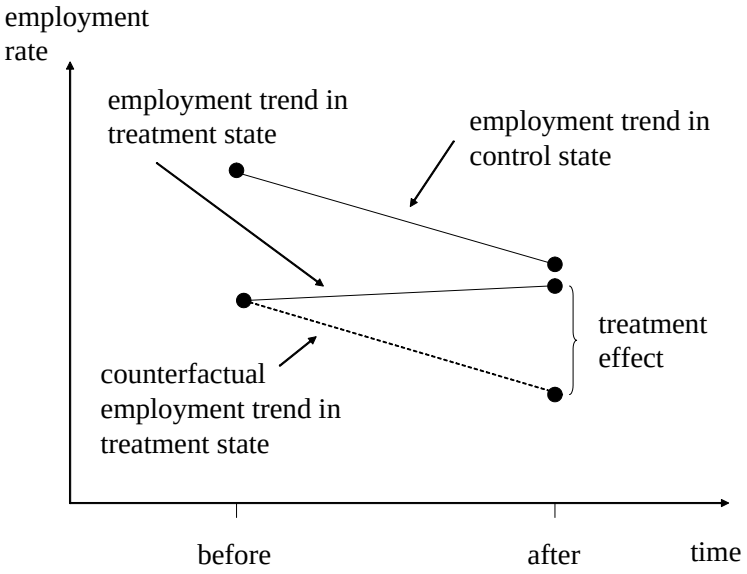

Figure 5.2.1: Causal effects in the differences-in-differences model

original minimum wage study, Card and Krueger (2000) obtained administrative payroll data for restaurants in New Jersey and Pennsylvania for a number of years. These data are shown here in Figure 5.2.2, similar to Figure 2 in their follow-up study. The vertical lines indicate the dates when their original surveys were conducted, and the third vertical line denotes the increase in the federal minimum wage to $4.75 in October 1996, which affected Pennsylvania but not New Jersey. These data give us an opportunity to look at a new minimum wage "experiment".

Like the original Card and Krueger survey, the administrative data show a slight decline in employment from February to November 1992 in Pennsylvania, and little change in New Jersey over the same period. However, the data also reveal fairly substantial year-to-year employment variation in other periods. These swings often seem to differ substantially in the two states. In particular, while employment levels in New Jersey and Pennsylvania were similar at the end of 1991, employment in Pennsylvania fell relative to employment in New Jersey over the next three years (especially in the 14-county group), mostly before the 1996 change in Federal minimum. So Pennsylvania may not provide a very good measure of counterfactual employment rates in New Jersey in the absence of a policy change, and vice versa.

A more encouraging example comes from Pischke (2007), who looks at the effect of school term length on student performance using variation generated by a sharp policy change in Germany. Until the 1960s, children in all German states except Bavaria started school in the Spring. Beginning in the 1966-67 school year, the Spring-starters moved to start school in the Fall. The transition to a Fall start required two short school years for affected cohorts, 24 weeks long instead of 37. Students in these cohorts effectively had their time in school compressed relative to cohorts on either side and relative to students in Bavaria, which

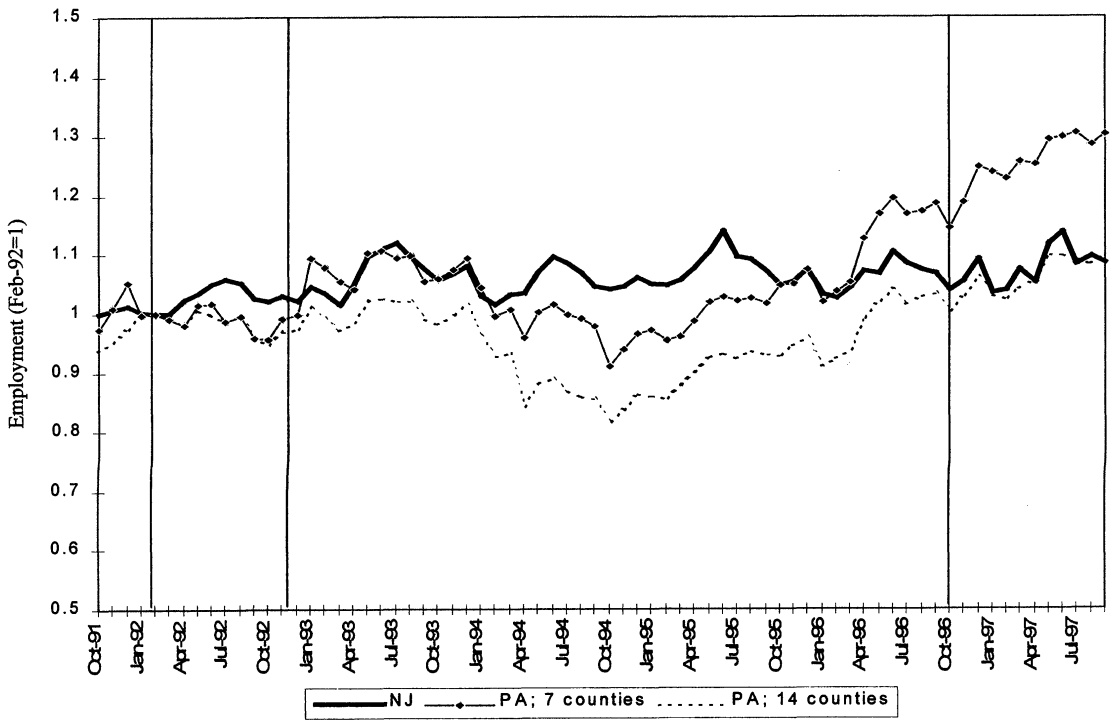

Figure 5.2.2: Employment in New Jersey and Pennsylvania fast-food restaurants, October 1991 to September 1997 (from Card and Krueger 2000). Vertical lines indicate dates of the original Card and Krueger (1994) survey and the October 1996 federal minimum-wage increase.

already had a Fall start.

Figure 5.2.3 plots the likelihood of grade repetition for the 1962-73 cohorts of 2nd graders in Bavaria and affected states (there are no repetition data for 1963-65). Repetition rates in Bavaria were reasonably flat from 1966 onwards at around 2.5%. Repetition rates are higher in the short-school-year states, at around 4 - 4.5% in 1962 and 1966, before the change in term length. But repetition rates jump up by about a percentage point for the two affected cohorts in these states, a bit more so for the second cohort than the first, before falling back to the baseline level. This graph provides strong visual evidence of treatment and control states with a common underlying trend, and a treatment effect that induces a sharp but transitory deviation from this trend. A shorter school year seems to have increased repetition rates for affected cohorts.

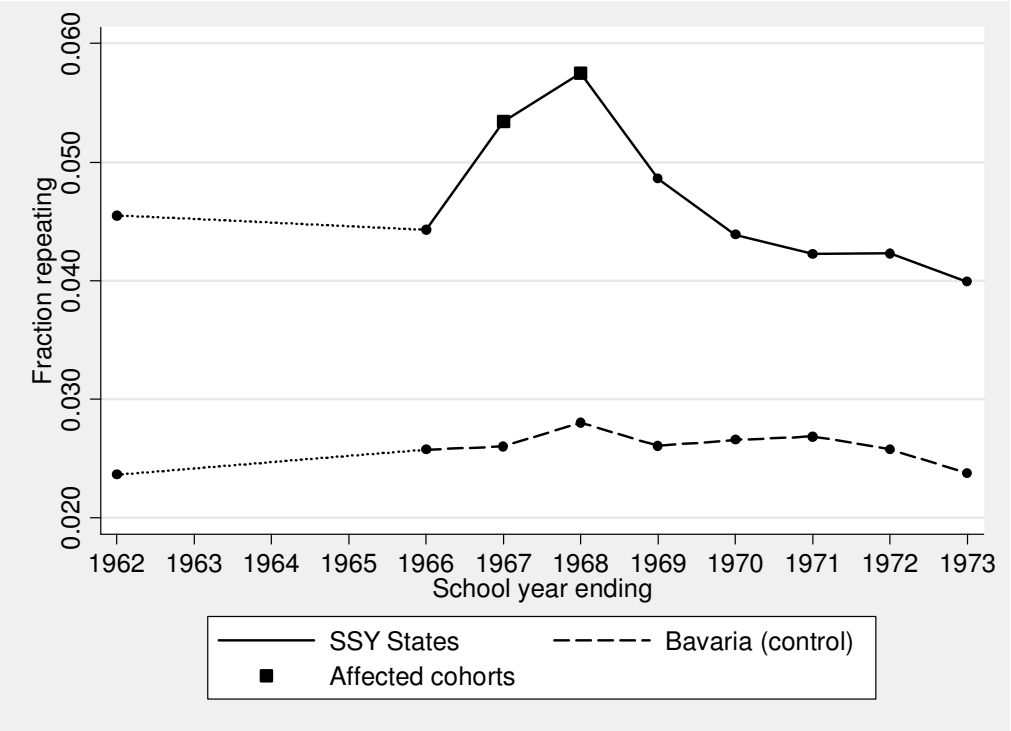

Figure 5.2.3: Average rates of grade repetition in second grade for treatment and control schools in Germany (from Pischke 2007). The data span a period before and after a change in term length for students outside of Bavaria.

5.2.1 Regression DD

As with the fixed effects model, we can use regression to estimate equations like (5.2.2). Let NJ_s be a dummy for restaurants in New Jersey and d_t be a time-dummy that switches on for observations obtained

in November (i.e., after the minimum wage change). Then

$$\text{Y}_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \beta(NJ_s \cdot d_t) + \varepsilon_{ist} \tag{5.2.3}$$

is the same as (5.2.2) where $NJ_s \cdot d_t = \text{D}_{st}$. In the language of Section 3.1.4, this model includes two main effects for state and year and an interaction term that marks observations from New Jersey in November. This is a saturated model since the conditional mean function $E(\text{Y}_{ist}|s,t)$ takes on four possible values and there are four parameters. The link between the parameters in the regression equation, (5.2.3), and those in the DD model for the conditional mean function, (5.2.2), is

$$\begin{aligned}
\alpha &= E(\text{Y}_{ist}|s = PA, t = Feb) = \gamma_{PA} + \lambda_{Feb} \\
\gamma &= E(\text{Y}_{ist}|s = NJ, t = Feb) - E(\text{Y}_{ist}|s = PA, t = Feb) = \gamma_{NJ} - \gamma_{PA} \\
\lambda &= E(\text{Y}_{ist}|s = PA, t = Nov) - E(\text{Y}_{ist}|s = PA, t = Feb) = \lambda_{Nov} - \lambda_{Feb} \\
\beta &= \{E(\text{Y}_{ist}|s = NJ, t = Nov) - E(\text{Y}_{ist}|s = NJ, t = Feb)\} \\
&\quad -\{E(\text{Y}_{ist}|s = PA, t = Nov) - E(\text{Y}_{ist}|s = PA, t = Feb)\}.
\end{aligned}$$

The regression formulation of the difference-in-difference model offers a convenient way to construct DD estimates and standard errors. It's also easy to add additional states or periods to the regression set-up. We might for example, add additional control states and pre-treatment periods to the New Jersey/Pennsylvania sample. The resulting generalization of (5.2.3) includes a dummy for each state and period but is otherwise unchanged.

A second advantage of regression-DD is that it facilitates empirical work with regressors other than switched-on/switched-off dummy variables. Instead of New Jersey and Pennsylvania in 1992, for example, we might look at all state minimum wages in the United States. Some of these are a little higher than the federal minimum (which covers everyone regardless of where they live), some are a lot higher, and some are the same. The minimum wage is therefore a variable with differing "treatment intensity" across states and over time. Moreover, in addition to statutory variation in state minima, the local importance of a minimum wage varies with average state wage levels. For example, the early-1990s Federal minimum of $4.25 was probably irrelevant in Connecticut - with high average wages - but a big deal in Mississippi.

Card (1992) exploits regional variation in the impact of the federal minimum wage. His approach is motivated by an equation like

$$\text{Y}_{ist} = \gamma_s + \lambda_t + \beta(\text{FA}_s \cdot d_t) + \varepsilon_{ist} \tag{5.2.4}$$

where the variable FA_s is a measure of the fraction of teenagers likely to be affected by a minimum wage increase in each state and d_t is a dummy for observations after 1990, when the federal minimum increased from $3.35 to $3.80. The FA_s variable measures the baseline (pre-increase) proportion of each state's teen

labor force earning less than $3.80.

As in the New Jersey/Pennsylvania study, Card (1992) works with data from two periods, before and after, in this case 1989 and 1992. But this study uses 51 states (including the District of Columbia), for a total of 102 state-year observations. Since there are no individual-level covariates in (5.2.4), this is the same as estimation with micro data (provided the group-level estimates are weighted by cell size). Note that $\text{FA}_s \cdot d_t$ is an interaction term, like $NJ_s \cdot d_t$ in (5.2.3), though here the interaction term takes on a distinct value for each observation in the data set. Finally, because Card (1992) analyzes data for only two periods, the reported estimates are from an equation in first-differences:

$$\Delta\bar{\text{Y}}_s = \lambda^* + \beta\text{FA}_s + \Delta\bar{\varepsilon}_s,$$

where $\Delta\bar{\text{Y}}_s$ is the change in average teen employment in state s and $\Delta\bar{\varepsilon}_s$ is the error term in the differenced equation.[8]

Table 5.2.2, based on Table 3 in Card (1992), shows that wages increased more in states where the minimum wage increase is likely to have had more bite (see the estimate of .15 in column 1). This is an important step in Card's analysis - it verifies the notion that the *fraction affected* variable is a good predictor of the wage changes induced by an increase in the federal minimum. Employment, on the other hand, seems largely unrelated to *fraction affected*, as can be seen in column 3. Thus, the results in Card (1992) are in line with the results from the New Jersey/Pennsylvania study.

Table 5.2.2: Regression-DD estimates of minimum wage effects on teens, 1989 to 1992

	Equations for Change in Mean Log Wage:		Equations for change in Teen Employment-Population Ratio:	
Explanatory Variable	(1)	(2)	(3)	(4)
1. Fraction of Affected Teens	0.15 (0.03)	.14 (0.04)	0.02 (0.03)	-.01 (0.03)
2. Change in Overall Emp./Pop. Ratio	–	0.46 (0.60)	–	1.24 (0.60)
3. R-squared	0.30	0.31	0.01	0.09

Notes: Adapted from Card (1992). The table reports estimates from a regression of the change in average teen employment by state on the fraction of teens affected by a change in the federal minimum wage in each state. Data are from the 1989 and 1992 CPS. Regressions are weighted by the CPS sample size by state and year.

Card's (1992) analysis illustrates a further advantage of regression-DD: it's easy to add additional covariates in this framework. For example, we might like to control for adult employment as a source of omitted

[8] Card weights estimates of (5.2.4) by the sample size used to construct averages for each state. Other specifications in the spirit of (5.2.4) put a normalized function of state and federal minimum wages on the right hand side instead of $\text{FA}_s \cdot d_t$. See, for example, Neumark and Wascher (1992), who work with the difference between state and federal minima, adjusted for minimum-wage coverage provisions, and normalized by state average hourly wages.

state-specific trends. In other words, we can model counterfactual employment in the absence of a change in the minimum wage as

$$E[\mathrm{Y}_{0ist}|s,t,\mathrm{X}_{st}] = \gamma_s + \lambda_t + \mathrm{X}'_{st}\delta.$$

where X_{st} is a vector of state-and-time-varying covariates, including adult employment (though this may not be kosher if adult employment also responds to the minimum wage change, in which case it's *bad control*; see Section 3.2.3). As it turns out, the addition of an adult employment control has little effect on Card's estimates, as can be seen in columns 2 and 4 in Table 5.2.2.

It's worth emphasizing the fact that Card (1992) analyzes state averages instead of individual data. He might have used a pooled multi-year sample of micro data from the CPS to estimate an equation like

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \beta(\mathrm{FA}_s \cdot d_t) + \mathrm{X}'_{ist}\delta + \varepsilon_{ist}, \tag{5.2.5}$$

where X_{ist} can include individual level characteristics such as race. The covariate vector might also include time-varying variables measured at the state level. Only the latter are likely to be a source of omitted variables bias, but individual-level controls can increase precision, a point we noted in Section 2.3. Inference is a little more complicated in a framework that combines of micro data on dependent variables with group-level regressors, however. The key issue is how best to adjust for possible group-level random effects, as we discuss in Chapter 8, below.

When the sample includes many years, the regression-DD model lends itself to a test for causality in the spirit of Granger (1969). The Granger idea is to see whether causes happen before consequences and not vice versa (though as we know from the epigram at the beginning of Chapter 4, this alone is not sufficient for causal inference). Suppose the policy variable of interest, D_{st}, changes at different times in different states. In this context, Granger causality testing means a check on whether, conditional on state and year effects, past D_{st} predicts Y_{ist} while future D_{st} does not. If D_{st} causes Y_{ist} but not vice versa, then leads should not matter in an equation like:

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \sum_{\tau=0}^{m} \beta_{-\tau}\mathrm{D}_{s,t-\tau} + \sum_{\tau=1}^{q} \beta_{+\tau}\mathrm{D}_{s,t+\tau}\mathrm{X}_{ist}\delta + \varepsilon_{ist}, \tag{5.2.6}$$

where the sums on the right-hand side allow for m lags $(\beta_{-1}, \beta_{-2}, ..., \beta_{-m})$ or post-treatment effects and q leads $(\beta_{+1}, \beta_{+1}, ..., \beta_{+q})$ or anticipatory effects. The pattern of lagged effects is usually of substantive interest as well. We might, for example, believe that causal effects should grow or fade as time passes.

Autor (2003) implements the Granger test in an investigation of the effect of employment protection on firms' use of temporary help. Employment protection is a type of labor law - promulgated by state legislatures or, more typically, through common law as made by state courts - that makes it harder to fire workers. As a rule, U.S. labor law allows "employment at will," which means that workers can be fired for

just cause or no cause, at the employer's whim. But some state courts have allowed a number of exceptions to the employment-at-will doctrine, leading to lawsuits for "unjust dismissal". Autor is interested in whether fear of employee lawsuits makes firms more likely to use temporary workers for tasks for which they would otherwise have increased their workforce. Temporary workers work for someone else besides the firm for which they are executing tasks. As a result, the firm using them cannot be sued for unjust dismissal when they let temporary workers go.

Autor's empirical strategy relates the employment of temporary workers in a state to dummy variables indicating state court rulings that allow exceptions to the employment-at-will doctrine. His regression-DD model includes both leads and lags, as in equation (5.2.6). The estimated leads and lags, running from two years ahead to 4 years behind, are plotted in Figure 5.2.4, a reproduction of Figure 3 from Autor (2003). The estimates show no effects in the two years before the courts adopted an exception, with sharply increasing effects on temporary employment in the first few years after the adoption, which then appear to flatten out with a permanently higher rate of temporary employment in affected states. This pattern seems consistent with a causal interpretation of Autor's results.

An alternative check on the DD identification strategy adds state-specific time trends to the regressors in X_{ist}. In other words, we estimate

$$\mathrm{Y}_{ist} = \gamma_{0s} + \gamma_{1st} + \lambda_t + \beta \mathrm{D}_{st} + \mathrm{X}'_{ist}\delta + \varepsilon_{ist}, \tag{5.2.7}$$

where γ_{0s} is a state-specific intercept as before and γ_{1s} is a state-specific trend coefficient multiplying the time-trend variable, t. This allows treatment and control states to follow different trends in a limited but potentially revealing way. It's heartening to find that the estimated effects of interest are unchanged by the inclusion of these trends, and discouraging otherwise. Note, however, that we need at least 3 periods to estimate a model with state-specific trends. Moreover, in practice, 3 periods is typically inadequate to pin down both the trends and the treatment effect. As a rule, DD estimation with state-specific trends is likely to be more robust and convincing when the pre-treatment data establish a clear trend that can be extrapolated into the post-treatment period.

In a study of the effect of labor regulation on businesses in Indian states, Besley and Burgess (2004)use state trends as a robustness check. Different states change regulatory regimes at different times, giving rise to a DD research design. As in Card (1992), the unit of observation in Besley and Burgess (2004) is a state-year average. Table 5.2.3 (based on Table IV in their paper) reproduces the key results.

The estimates in column 1, from a regression-DD model without state-specific trends, suggest that labor regulation leads to lower output per capita. The models used to construct the estimates in columns 2 and 3 add time-varying state-specific covariates like government expenditure per capita and state population. This is in the spirit of Card's (1992) addition of state-level adult employment rates as a control in the minimum

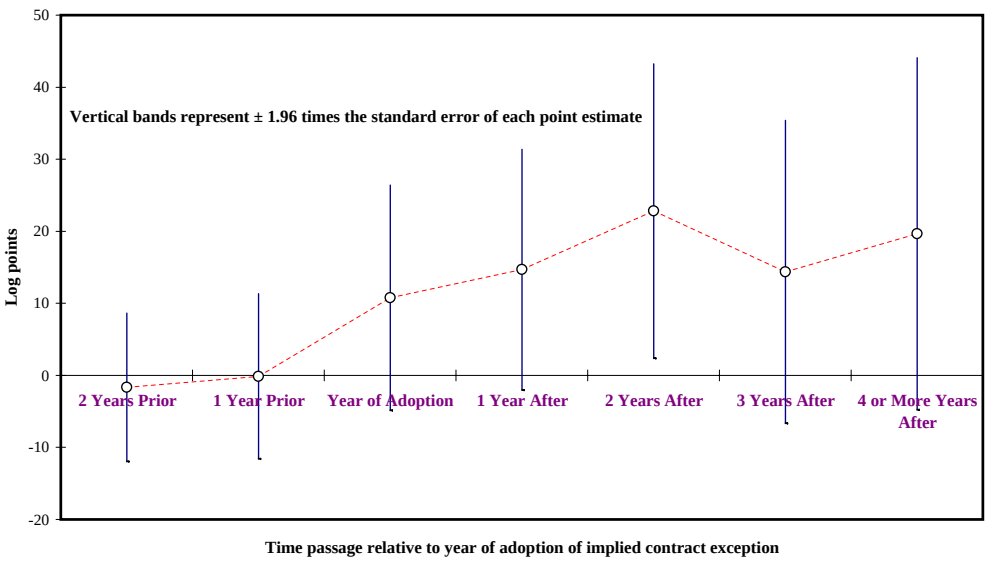

Figure 5.2.4: Estimated impact of state courts' adoption of an implied-contract exception to the employment-at-will doctrine on use of temporary workers (from Autor 2003). The dependent variable is the log of state temporary help employment in 1979 - 1995. Estimates are from a model that allows for effects before, during, and after adoption.

Table 5.2.3: Effect of labor regulation on the performance of firms in Indian states

	(1)	(2)	(3)	(4)
Labor regulation (lagged)	-0.186	-0.185	-0.104	0.0002
	(.0641)	(.0507)	(.039)	(.02)
Log development		0.240	0.184	0.241
expenditure per capita		(.1277)	(.1187)	(.1057)
Log installed electricity		0.089	0.082	0.023
capacity per capita		(.0605)	(.0543)	(.0333)
Log state population		0.720	0.310	-1.419
		(.96)	(1.1923)	(2.3262)
Congress majority			-0.0009	0.020
			(.01)	(.0096)
Hard left majority			-0.050	-0.007
			(.0168)	(.0091)
Janata majority			0.008	-0.020
			(.0235)	(.0333)
Regional majority			0.006	0.026
			(.0086)	(.0234)
State-specific trends	NO	NO	NO	YES
Adjusted R-squared	0.93	0.93	0.94	0.95

Notes: Adapted from Besley and Burgess (2004), Table IV. The table reports regression-DD estimates of the effects of labor regulation on productivity. The dependent variable is log manufacturing output per capita. All models include state and year effects. Robust standard errors clustered at the state level are reported in parentheses. State amendments to the Industrial Disputes Act are coded 1=pro-worker, 0 = neutral, -1 = pro-employer and then cumulated over the period to generate the labor regulation measure. Log of installed electrical capacity is measured in kilowatts, and log development expenditure is real per capita state spending on social and economic services. Congress, hard left, Janata, and regional majority are counts of the number of years for which these political groupings held a majority of the seats in the state legislatures. The data are for the sixteen main states for the period 1958-1992. There are 552 observations.

wage study. The addition of controls affects the Besley and Burgess estimates little. But the addition of state-specific trends kills the labor-regulation effect, as can be seen in column 4. Apparently, labor regulation in India increases in states where output is declining anyway. Control for this trend therefore drives the estimated regulation effect to zero.

Picking Controls

We've labeled the two dimensions in the DD set-up "states" and "time" because this is the archetypical DD example in applied econometrics. But the DD idea is much more general. Instead of states, the subscript s might denote demographic groups, some of which are affected by a policy and others are not. For example, Kugler, Jimeno, and Hernanz (2005) look at the effects of age-specific employment protection

policies in Spain. Likewise, instead of time, we might group data by cohort or other types of characteristics. An example is Angrist and Evans (1999), who study the effect of changes in state abortion laws on teen pregnancy using variation by state and year of birth. Implicitly, however, DD designs always set up an implicit treatment-control comparison. The question of whether this comparison is a good one deserves careful consideration.

One potential pitfall in this context arises when the composition of the implicit treatment and control groups changes as a result of treatment. Going back to a design based on state/time comparisons, suppose we're interested in the effects of the generosity of public assistance on labor supply. Historically, U.S. states have offered widely-varying welfare payments to poor unmarried mothers. Labor economists have long been interested in the effects of such income maintenance policies - how much of an increase in living standards they facilitate, and whether they make work less attractive (see, e.g., Meyer and Rosenbaum, 2001, for a recent study). A concern here, emphasized in a review of research on welfare by Moffitt (1992), is that poor people who would in any case have weak labor force attachment might move to states with more generous welfare benefits. In a DD research design, this sort of program-induced migration tends to make generous welfare programs look worse for labor supply than they really are.

Migration problems can usually be fixed if we know where an individual starts out. Say we know state of residence in the period before treatment, or state of birth. State of birth or previous state of residence are unchanged by the treatment but still highly correlated with current state of residence. The problem of migration is therefore eliminated in comparisons using these dimensions instead of state of residence. This introduces a new problem, however, which is that individuals who do move are incorrectly located. In practice, however, this problem is easily addressed with the IV methods discussed in chapter 4 (state of birth or previous residence is used to construct instruments for current location).

A modification of the two-by-two DD set-up uses higher-order contrasts to draw causal inferences. An example is the extension of Medicaid coverage in the U.S. studied by Yelowitz (1995). Eligibility for Medicaid, the massive U.S. health insurance program for the poor, was once tied to eligibility for AFDC, a large cash welfare program. At various times in the 1980s, however, some states extended Medicaid coverage to children in families ineligible for AFDC. Yelowitz was interested in how this expansion affected, among other things, mothers' labor force participation and earnings.

In addition to state and time, children's age provides a third dimension in which Medicaid policy varies. Yelowitz exploits this variation by estimating

$$\mathrm{Y}_{iast} = \gamma_{st} + \lambda_{at} + \theta_{as} + \beta \mathrm{D}_{ast} + \mathrm{X}_{iast}\delta + \varepsilon_{iast},$$

where s index states, t indexes time, and a is the age of the youngest child in a family. This model provides full non-parametric control for state-specific time effects that are common across age groups (γ_{st}), time-varying

age effects (λ_{at}), and state-specific age effects (θ_{as}). The regressor of interest, D_{ast}, indicates children in affected age groups in states and periods where coverage is provided. This triple-differences model may generate a more convincing set of results than a traditional DD analysis that exploits differences by state and time alone.

5.3 Fixed Effects versus Lagged Dependent Variables

Fixed effects and differences-in-differences estimators are based on the presumption of time-invariant (or group-invariant) omitted variables. Suppose, for example, we are interested in the effects of participation in a subsidized training program, as in the Dehejia and Wahba (1999) and Lalonde (1986) studies discussed in section (3.3.3). The key identifying assumption motivating fixed effects estimation in this case is

$$E(\mathrm{Y}_{0it}|\alpha_i, \mathrm{X}_{it}, \mathrm{D}_{it}) = E(\mathrm{Y}_{0it}|\alpha_i, \mathrm{X}_{it}), \tag{5.3.1}$$

where α_i is an unobserved personal characteristic that determines, along with covariates, X_{it}, whether individual i gets training. To be concrete, α_i might be a measure of vocational skills, though a strike against the fixed-effects setup is the fact that the exact nature of the unobserved variables typically remains somewhat mysterious. In any case, coupled with a linear model for $E(\mathrm{Y}_{0it}|\alpha_i,\mathrm{X}_{it})$, assumption (5.3.1) leads to simple estimation strategies involving differences or deviations from means.

For many causal questions, the notion that the most important omitted variables are time-invariant doesn't seem plausible. The evaluation of training programs is a case in point. It seems likely that people looking to improve their labor market options by participating in a government-sponsored training program have suffered some kind of setback. Many training programs explicitly target people who have suffered a recent setback, e.g., men who recently lost their jobs. Consistent with this, Ashenfelter (1978) and Ashenfelter and Card (1985) find that training participants typically have earnings histories that exhibit a pre-program dip. Past earnings is a time-varying confounder that cannot be subsumed in a time-invariant variable like α_i.

The distinctive earnings histories of trainees motivates an estimation strategy that controls for past earnings directly and dispenses with the fixed effects. To be precise, instead of (5.3.1), we might base causal inference on the conditional independence assumption,

$$E(\mathrm{Y}_{0it}|\mathrm{Y}_{it-h}, \mathrm{X}_{it}, \mathrm{D}_{it}) = E(\mathrm{Y}_{0it}|\mathrm{Y}_{it-h}, \mathrm{X}_{it}). \tag{5.3.2}$$

This is like saying that what makes trainees special is their earnings h periods ago. We can then use panel data to estimate

$$\mathrm{Y}_{it} = \alpha + \theta \mathrm{Y}_{it-h} + \lambda_t + \beta \mathrm{D}_{it} + \mathrm{X}_{it}\delta + \varepsilon_{it}, \tag{5.3.3}$$

where the causal effect of training is β. To make this more general, Y_{it-h} can be a vector including lagged earnings for multiple periods.[9]

Applied researchers using panel data are often faced with the challenge of choosing between fixed-effects and lagged-dependent variables models, i.e., between causal inferences based on (5.3.1) and (5.3.2). One solution to this dilemma is to work with a model that includes both lagged dependent variables and unobserved individual effects. In other words, identification might be based on a weaker conditional independence assumption:

$$E(\mathrm{Y}_{0it}|a_i, \mathrm{Y}_{it-h}, \mathrm{X}_{it}, \mathrm{D}_{it}) = E(\mathrm{Y}_{0it}|\alpha_i, \mathrm{Y}_{it-h}, \mathrm{X}_{it}), \tag{5.3.4}$$

which requires conditioning on both α_i and Y_{it-h}. We can then try to estimate causal effects using a specification like

$$\mathrm{Y}_{it} = \alpha_i + \theta \mathrm{Y}_{it-h} + \lambda_t + \beta \mathrm{D}_{it} + \mathrm{X}_{it}\delta + \upsilon_{it}. \tag{5.3.5}$$

Unfortunately, the conditions for consistent estimation of β in equation (5.3.5) are much more demanding than those required with fixed effects or lagged dependent variables alone. This can be seen in a simple example where the lagged dependent variable is Y_{it-1}. We kill the fixed effect by differencing, which produces

$$\Delta \mathrm{Y}_{it} = \theta \Delta \mathrm{Y}_{it-1} + \Delta\lambda_t + \beta \Delta \mathrm{D}_{it} + \Delta \mathrm{X}_{it}\delta + \Delta\upsilon_{it}. \tag{5.3.6}$$

The problem here is that the differenced residual, $\Delta\upsilon_{it}$, is necessarily correlated with the lagged dependent variable, $\Delta \mathrm{Y}_{it-1}$, because both are a function of υ_{it-1}. Consequently, OLS estimates of (5.3.6) are not consistent for the parameters in (5.3.5), a problem first noted by Nickell (1981). This problem can be solved, though the solution requires strong assumptions. The easiest solution is to use Y_{it-2} as an instrument for $\Delta \mathrm{Y}_{it-1}$ in (5.3.6).[10] But this requires that Y_{it-2} be uncorrelated with the differenced residuals, $\Delta\upsilon_{it}$. This seems unlikely since residuals are the part of earnings left over after accounting for covariates. Most people's earnings are highly correlated from one year to the next, so that past earnings are an excellent predictor of future earnings and earnings growth . If υ_{it} is serially correlated, there may be no consistent estimator for (5.3.6). (Note also that the IV strategy using Y_{it-2} as an instrument requires at least three periods to obtain data for t, $t-1$, and $t-2$).

Given the difficulties that arise when trying to estimate (5.3.6), we might ask whether the distinction between fixed effects and lagged dependent variables matters. The answer, unfortunately, is yes. The fixed-effects and lagged dependent variables models are not nested, which means we cannot hope to estimate

[9] Abadie, Diamond, and Hainmueller (2007) develop a semiparametric version of the lagged-dependent variables model, more flexible than the traditional regression setup. As with our regression setup, the key assumption in this model is conditional independence of potential outcomes conditional on lagged earnings, i.e., assumption (5.3.2).

[10] See Holtz-Eakin, Newey and Rosen (1988), Arellano and Bond (1991), Blundell and Bond (1998) for details and examples.

one and get the other as a special case if need be. Only the more general and harder-to-identify model, (5.3.5), nests both fixed effects and lagged dependent variables.[11].

So what's an applied guy to do? One answer, as always, is to check the robustness of your findings using alternative identifying assumptions. That means that you would like to find broadly similar results using both models. Fixed effects and lagged dependent variables estimates also have a useful bracketing property. The appendix to this chapter shows that if (5.3.2) is correct, but you mistakenly use fixed effects, estimates of a positive treatment effect will tend to be too big. On the other hand, if (5.3.1) is correct and you mistakenly estimate an equation with lagged outcomes like (5.3.3), estimates of a positive treatment effect will tend to be too small. You can therefore think of fixed effects and lagged dependent variables as bounding the causal effect you are after. Guryan (2004) illustrates this sort of reasoning in a study estimating the effects of court-ordered busing on Black high school graduation rates.

5.4 Appendix: More on fixed effects and lagged dependent variables

To simplify, we ignore covariates and year effects and assume there are only two periods, with treatment equal to zero for everyone in the first period (the punch line is the same in a more general setup). The causal effect of interest, β, is positive. Suppose first that treatment is correlated with an unobserved individual effect, a_i, and that outcomes can be described by

$$\mathrm{Y}_{it} = a_i + \beta \mathrm{D}_{it} + \varepsilon_{it}. \tag{5.4.1}$$

where ε_{it} is serially uncorrelated, and uncorrelated with a_i and D_{it}. We also have

$$\mathrm{Y}_{it-1} = a_i + \varepsilon_{it-1},$$

where a_i and ε_{it-1} are uncorrelated. You mistakenly estimate the effect of D_{it} in a model that controls for Y_{it-1} but ignores fixed effects. The resulting estimator has probability limit $\frac{Cov(\mathrm{Y}_{it},\tilde{\mathrm{D}}_{it})}{V(\mathrm{D}_{it})}$, where $\tilde{\mathrm{D}}_{it} = \mathrm{D}_{it} - \gamma \mathrm{Y}_{it-1}$ is the residual from a regression of D_{it} on Y_{it-1}.

[11] In particular, setting $\theta = 1$ in (5.3.3) does not produce the fixed-effects model as a special case of the lagged dependent variables model. Instead we get

$$\Delta \mathrm{Y}_{it} = \alpha + \lambda_t + \beta \mathrm{D}_{it} + \mathrm{X}_{it}\delta + \varepsilon_{it}$$

i.e., a differenced dependent variable with regressors in levels. This is not the model with first differences on both the right and left side needed to kill the fixed effect.

Now substitute $a_i = \mathrm{Y}_{it-1} - \varepsilon_{it-1}$ in (5.4.1) to get

$$\mathrm{Y}_{it} = \mathrm{Y}_{it-1} + \beta \mathrm{D}_{it} + \varepsilon_{it} - \varepsilon_{it-1}.$$

From here, we get

$$\frac{Cov(\mathrm{Y}_{it}, \tilde{\mathrm{D}}_{it})}{V(\tilde{\mathrm{D}}_{it})} = \beta - \frac{Cov(\varepsilon_{it-1}, \tilde{\mathrm{D}}_{it})}{V(\tilde{\mathrm{D}}_{it})} = \beta - \frac{Cov(\varepsilon_{it-1}, \mathrm{D}_{it} - \gamma \mathrm{Y}_{it-1})}{V(\tilde{\mathrm{D}}_{it})} = \beta + \frac{\gamma \sigma_{\varepsilon}^2}{V(\tilde{\mathrm{D}}_{it})}.$$

where σ_{ε}^2 is the variance of ε_{it-1}. Since trainees have low Y_{it-1}, $\gamma < 0$ and the resulting estimate of β is too small.

Suppose instead that treatment is determined by low Y_{it-1}. The correct specification is a simplified version of (5.3.3), say

$$\mathrm{Y}_{it} = \alpha + \theta \mathrm{Y}_{it-1} + \beta \mathrm{D}_{it} + \varepsilon_{it}, \tag{5.4.2}$$

where ε_{it} is serially uncorrelated. You mistakenly estimate a first-differenced equation in an effort to kill fixed effects. This ignores lagged dependent variables. In this simple example, where $\mathrm{D}_{it-1} = 0$ for everyone, the first-differenced estimator has probability limit

$$\frac{Cov(\mathrm{Y}_{it} - \mathrm{Y}_{it-1}, \mathrm{D}_{it} - \mathrm{D}_{it-1})}{V(\mathrm{D}_{it} - \mathrm{D}_{it-1})} = \frac{Cov(\mathrm{Y}_{it} - \mathrm{Y}_{it-1}, \mathrm{D}_{it})}{V(\mathrm{D}_{it})}. \tag{5.4.3}$$

Subtracting Y_{it-1} from both sides of (5.4.2), we have

$$\mathrm{Y}_{it} - \mathrm{Y}_{it-1} = \alpha + (\theta - 1)\mathrm{Y}_{it-1} + \beta \mathrm{D}_{it} + \varepsilon_{it}.$$

Substituting this in (4.2.2), the inappropriately differenced model yields

$$\frac{Cov(\mathrm{Y}_{it} - \mathrm{Y}_{it-1}, \mathrm{D}_{it})}{V(\mathrm{D}_{it})} = \beta + (\theta - 1)\left[\frac{Cov(\mathrm{Y}_{it-1}, \mathrm{D}_{it})}{V(\mathrm{D}_{it})}\right].$$

In general, we think θ is a number between zero and one, otherwise Y_{it} is non-stationary (i.e., an explosive time series process). Therefore, since trainees have low $\mathrm{Y}_{it-1,}$ the estimate of β in first differences is too big.

Part III

Extensions

Chapter 6

Getting a Little Jumpy: Regression Discontinuity Designs

> But when you start exercising those rules, all sorts of processes start to happen and you start to find out all sorts of stuff about people . . . Its just a way of thinking about a problem, which lets the shape of the problem begin to emerge. The more rules, the tinier the rules, the more arbitrary they are, the better.
>
> Douglas Adams, *Mostly Harmless* (1995)

Regression discontinuity (RD) research designs exploit precise knowledge of the rules determining treatment. RD identification is based on the idea that in a highly rule-based world, some rules are arbitrary and therefore provide good experiments. RD comes in two styles, fuzzy and sharp. The sharp design can be seen as a selection-on-observables story. The fuzzy design leads to an instrumental-variables-type setup.

6.1 Sharp RD

Sharp RD is used when treatment status is a deterministic and discontinuous function of a covariate, x_i. Suppose, for example, that

$$\text{D}_i = \begin{cases} 1 & \text{if } x_i \geq x_0 \\ 0 & \text{if } x_i < x_0 \end{cases}. \tag{6.1.1}$$

where x_0 is a known threshold or cutoff. This assignment mechanism is a deterministic function of x_i because once we know x_i we know D_i. It's a discontinuous function because no matter how close x_i gets to x_0, treatment is unchanged until $x_i = x_0$.

This may seem a little abstract, so here is an example. American high school students are awarded National Merit Scholarship Awards on the basis of PSAT scores, a test taken by most college-bound high

school juniors, especially those who will later take the SAT. The question that motivated the first discussions of RD is whether students who win these awards are more likely to finish college (Thistlewaithe and Campbell, 1960; Campbell, 1969). Sharp RD compares the college completion rates of students with PSAT scores just above and just below the National Merit Award thresholds. In general, we might expect students with higher PSAT scores to be more likely to finish college, but this effect can be controlled by fitting a regression to the relationship between college completion and PSAT scores, at least in the neighborhood of the award cutoff. In this example, jumps in the relationship between PSAT scores and college attendance in the neighborhood of the award threshold are taken as evidence of a treatment effect. It is this jump in regression lines that gives RD its name.[1]

An interesting and important feature of RD, highlighted in a recent survey of RD by Imbens and Lemieux (2008), is that there is *no* value of x_i at which we get to observe both treatment and control observations. Unlike full-covariate matching strategies, which are based on treatment-control comparisons conditional on covariate values where there is some overlap, the validity of RD turns on our willingness to extrapolate across covariate values, at least in a neighborhood of the discontinuity. This is one reason why sharp RD is usually seen as distinct from other control strategies. For this same reason, we typically cannot afford to be as agnostic about regression functional form in the RD world as in the world of Chapter 3.

Figure 6.1.1 illustrates a hypothetical RD scenario where those with $x_i \geq 0.5$ are treated. In Panel A, the trend relationship between Y_i and x_i is linear, while in Panel B, it's nonlinear. In both cases, there is a discontinuity in the relation between $E[\text{Y}_{0i}|x_i]$ and x_i around the point x_0.

A simple model formalizes the RD idea. Suppose that in addition to the assignment mechanism, (6.1.1), potential outcomes can be described by a linear, constant-effects model

$$\begin{aligned} E[\text{Y}_{0i}|x_i] &= \alpha + \beta x_i \\ \text{Y}_{1i} &= \text{Y}_{0i} + \rho \end{aligned}$$

This leads to the regression,

$$\text{Y}_i = \alpha + \beta x_i + \rho \text{D}_i + \eta_i, \tag{6.1.2}$$

where ρ is the causal effect of interest. The key difference between this regression and others we've used to estimate treatment effects (e.g., in Chapter 3) is that D_i, the regressor of interest, is not only correlated with x_i, it is a deterministic function of x_i. RD captures causal effects by distinguishing the nonlinear and discontinuous function, $1(x_i \geq x_0)$, from the smooth and (in this case) linear function, x_i.

[1] The basic structure of RD designs appears to have emerged simultaneously in a number of disciplines but has only recently become important in applied econometrics. Cook (2008) gives an intellectual history. In a recent paper using Lalonde (1986) style within-study comparisons, Cook and Wong (2008) find that RD generally does a good job of reproducing the results from randomized trials.

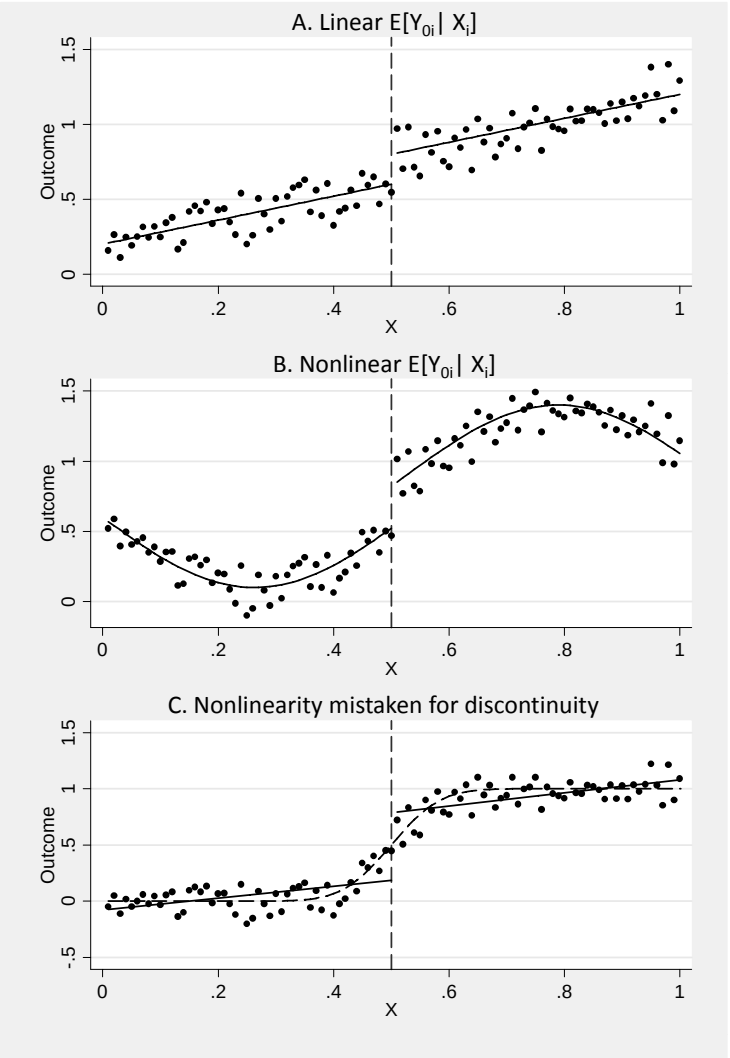

Figure 6.1.1: The sharp regression discontinuity design

But what if the trend relation, $E[\text{Y}_{0i}|x_i]$, is nonlinear? To be precise, suppose that $E[\text{Y}_{0i}|x_i] = f(x_i)$ for some reasonably smooth function, $f(x_i)$. Panel B in Figure 6.1.1 suggests there is still hope even in this more general case. Now we can construct RD estimates by fitting

$$\text{Y}_i = f(x_i) + \rho \text{D}_i + \eta_i, \tag{6.1.3}$$

where again, $\text{D}_i = 1(x_i \geq x_0)$ is discontinuous in x_i at x_0. As long as $f(x_i)$ is continuous in a neighborhood of x_0, it should be possible to estimate a model like (6.1.3), even with a flexible functional form for $f(x_i)$. For example, modeling $f(x_i)$ with a p^{th}-order polynomial, RD estimates can be constructed from the regression

$$\text{Y}_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p + \rho \text{D}_i + \eta_i. \tag{6.1.4}$$

A generalization of RD based on (6.1.4) allows different trend functions for $E[\text{Y}_{0i}|x_i]$ and $E[\text{Y}_{1i}|x_i]$. Modeling both of these CEFs with p^{th}-order polynomials, we have

$$\begin{aligned} E[\text{Y}_{0i}|x_i] &= f_0(x_i) = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + ... + \beta_{0p}\tilde{x}_i^p \\ E[\text{Y}_{1i}|x_i] &= f_1(x_i) = \alpha + \rho + \beta_{11}\tilde{x}_i + \beta_{12}\tilde{x}_i{}^2 + ... + \beta_{1p}\tilde{x}_i{}^p, \end{aligned}$$

where $\tilde{x}_i \equiv x_i - x_0$. Centering x_i at x_0 is just a normalization; it ensures that the treatment effect at $x_i = x_0$ is still the coefficient on D_i in the regression model with interactions.

To derive a regression model that can be used to estimate the effects interest in this case, we use the fact that D_i is a deterministic function of x_i to write

$$E[\text{Y}_i|x_i] = E[\text{Y}_{0i}|x_i] + E[\text{Y}_{1i} - \text{Y}_{0i}|x_i]\text{D}_i.$$

Substituting polynomials for conditional expectations, we then have

$$\begin{aligned} \text{Y}_i &= \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + ... + \beta_{0p}\tilde{x}_i^p \\ &\quad + \rho \text{D}_i + \beta_1^* \text{D}_i\tilde{x}_i + \beta_2^* \text{D}_i\tilde{x}_i{}^2 + ... + \beta_p^* \text{D}_i\tilde{x}_i{}^p + \eta_i, \end{aligned} \tag{6.1.6}$$

where $\beta_1^* = \beta_{11} - \beta_{01}$, $\beta_2^* = \beta_{12} - \beta_{02}$, and $\beta_p^* = \beta_{1p} - \beta_{0p}$ and the error term, η_i, is the CEF residual.

Equation (6.1.4) is a special case of (6.1.6) where $\beta_1^* = \beta_2^* = \beta_p^* = 0$. In the more general model, the treatment effect at $x_i - x_0 = c > 0$ is $\rho + \beta_1^* c + \beta_2^* c^2 + ... + \beta_p^* c^p$, while the treatment effect at x_0 is ρ. The model with interactions has the attraction that it imposes no restrictions on the underlying conditional mean functions But in our experience, RD estimates of ρ based on the simpler model, (6.1.4), usually turn out to be similar to those based on (6.1.6).

The validity of RD estimates based on (6.1.4) or (6.1.6) turns on whether polynomial models provide an adequate description of $E[\mathrm{Y}_{0i}|\mathrm{X}_i]$. If not, then what looks like a jump due to treatment might simply be an unaccounted-for nonlinearity in the counterfactual conditional mean function. This possibility is illustrated in Panel C of Figure 6.1.1, which shows how a sharp turn in $E[\mathrm{Y}_{0i}|x_i]$ might be mistaken for a jump from one regression line to another. To reduce the likelihood of such mistakes, we can look only at data in a neighborhood around the discontinuity, say the interval $[x_0 - \delta, x_0 + \delta]$ for some small number δ. Then we have

$$\begin{aligned} E\left[\mathrm{Y}_i|x_0 - \delta < x_i < x_0\right] &\simeq E[\mathrm{Y}_{0i}|x_i = x_0] \\ E\left[\mathrm{Y}_i|x_0 < x_i < x_0 + \delta\right] &\simeq E[\mathrm{Y}_{1i}|x_i = x_0], \end{aligned}$$

so that

$$\lim_{\delta \to 0} E\left[\mathrm{Y}_i|x_0 < x_i < x_0 + \delta\right] - E\left[\mathrm{Y}_i|x_0 - \delta < x_i < x_0\right] = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|x_i = x_0]. \tag{6.1.7}$$

In other words, comparisons of average outcomes in a small enough neighborhood to the left and right of x_0 should provide an estimate of the treatment effect that does not depend on the correct specification of a model for $E[\mathrm{Y}_{0i}|x_i]$. Moreover, the validity of this nonparametric estimation strategy does not turn on the constant effects assumption, $\mathrm{Y}_{1i} - \mathrm{Y}_{0i} = \rho$; the estimand in (6.1.7) is the average causal effect, $E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|x_i = x_0]$.

The nonparametric approach to RD requires good estimates of the mean of Y_i in small neighborhoods to the right and left of x_0. Obtaining such estimates is tricky. The first problem is that working in a small neighborhood of the cutoff means that you don't have much data. Also, the sample average is biased for the population average in the neighborhood of a boundary (in this case, x_0). Solutions to these problems include the use of a non-parametric version of regression called local linear regression (Hahn, Todd, and van der Klaauw, 2001) and the partial-linear and local-polynomial estimators developed by Porter (2003). Local linear regression amounts to weighted least squares estimation of an equation like (6.1.6), with linear terms only and more weight given to points close to the cutoff.

Sophisticated nonparametric RD methods have not yet found wide application in empirical practice; most applied RD work is still parametric. But the idea of focusing on observations near the cutoff value - what Angrist and Lavy (1999) call a "discontinuity sample" - suggests a valuable robustness check: Although RD estimates get less precise as the window used to select a discontinuity sample gets smaller, the number of polynomial terms needed to model $f(x_i)$ should go down. Hopefully, as you zero in on x_0 with fewer and fewer controls, the estimated effect of D_i remains stable.[2] A second important check looks at the behavior of

[2] Hoxby (2000) also uses this idea to check RD estimates of class size effects. A fully nonparametric approach requires data-driven rules for selection of the width of the discontinuity-sample window, also known as "bandwidth". The bandwidth must shrink with the sample size at a rate sufficiently slow so as to ensure consistent estimation of the underlying conditional mean functions. See Imbens and Lemieux (2007) for details. We prefer to think of estimation using (6.1.4) or (6.1.6) as essentially parametric: in any given sample, the estimates are only as good as the model for $E[\mathrm{Y}_{0i}|x_i]$ that you happen to be

pre-treatment variables near the discontinuity. Since pre-treatment variables are unaffected by treatment, there should be no jump in the CEF of these variables at x_0.

Lee's (2008) study of the effect of party incumbency on re-election probabilities illustrates the sharp RD design. Lee is interested in whether the Democratic candidate for a seat in the U.S. House of Representatives has an advantage if his party won the seat last time. The widely-noted success of House incumbents raises the question of whether representatives use the privileges and resources of their office to gain advantage for themselves or their parties. This conjecture sounds plausible, but the success of incumbents need not reflect a real electoral advantage. Incumbents - by definition, candidates and parties who have shown they can win - may simply be better at satisfying voters or getting the vote out.

To capture the causal effect of incumbency, Lee looks at the likelihood a Democratic candidate wins as a function of relative vote shares in the previous election. Specifically, he exploits the fact that an election winner is determined by $\text{D}_i = 1(x_i \geq .0)$, where x_i is the *vote share margin of victory* (e.g., the difference between the Democratic and Republican vote shares when these are the two largest parties). Note that, because D_i is a deterministic function of x_i, there are no confounding variables other than x_i. This is a signal feature of the RD setup.

Figure 6.1.2a, from Lee (2008), shows the sharp RD design in action. This figure plots the probability a Democrat wins against the difference between Democratic and Republican vote shares in the previous election. The dots in the figure are local averages (the average win rate in non-overlapping windows of share margins that are .005 wide); the lines in the figure are fitted values from a parametric model with a discontinuity at zero.[3] The probability of a democratic win is an increasing function of past vote share. The most important feature of the plot is the dramatic jump in win rates at the 0 percent mark, the point where a Democratic candidate gets more votes. Based on the size of the jump, incumbency appears to raise party re-election probabilities by about 40 percentage points.

Figure 6.1.2b checks the sharp RD identification assumptions by looking at Democratic victories *before* the last election. Democratic win rates in older elections should be unrelated to the cutoff in the last election, a specification check that works out well and increases our confidence in the RD design in this case. Lee's investigation of pre-treatment victories is a version of the idea that covariates should be balanced by treatment status in a (quasi-) randomized trial. A related check examines the density of x_i around the discontinuity, looking for bunching in the distribution of x_i near x_0. The concern here is that individuals with a stake in D_i might try to manipulate x_i near the cutoff, in which case observations on either side may not be comparable (McCrary 2008 proposes a formal test for this). Until recently, we would have said this is unlikely in election studies like Lee's. But the recount in Florida after the 2000 presidential election suggests we probably should worry about manipulable vote shares when U.S. elections are close.

---

using. Promises about how you might change the model if you had more data should be irrelevant.

[3] The fitted values in this figure are from a Logit model for the probability of winning as a function of the cutoff indicator $\text{D}_i = 1(x_i \geq 0)$, a 4^{th}-order polynomial in x_i, and interactions between the polynomial terms and D_i.

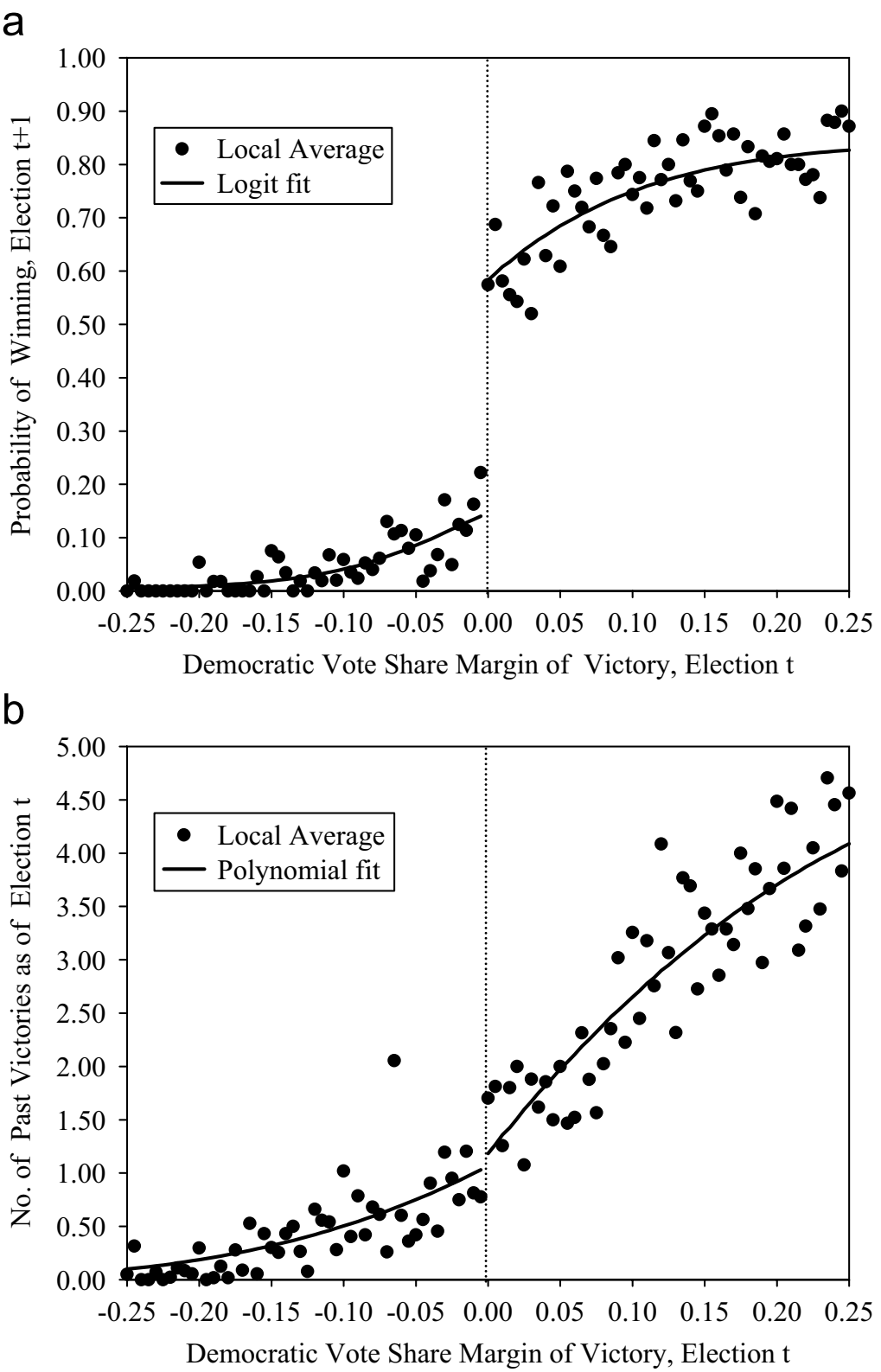

Figure 6.1.2: Probability of winning an election by past and future vote share (from Lee, 2008). (a) Candidate's probability of winning election $t+1$, by margin of victory in election t: local averages and parametric fit. (b) Candidate's accumulated number of past election victories, by margin of victory in election t: local averages and parametric fit.

6.2 Fuzzy RD is IV

Fuzzy RD exploits discontinuities in the *probability or expected value* of treatment conditional on a covariate. The result is a research design where the discontinuity becomes an instrumental variable for treatment status instead of deterministically switching treatment on or off. To see how this works, let D_i denote the treatment as before, though here D_i is no longer deterministically related to the threshold-crossing rule, $x_i \geq x_0$. Rather, there is a jump in the probability of treatment at x_0, so that

$$P[\text{D}_i = 1|x_i] = \begin{cases} g_0(x_i) & \text{if } x_i \geq x_0 \\ g_{1(}x_i) & \text{if } x_i < x_0 \end{cases}, \text{ where } g_1(x_0) \neq g_0(x_0).$$

The functions $g_0(x_i)$ and $g_1(x_i)$ can be anything as long as they differ (and the more the better) at x_0. We'll assume $g_1(x_0) > g_0(x_0)$, so $x_i \geq x_0$ makes treatment more likely. We can write the relation between the probability of treatment and x_i as

$$E\left[\text{D}_i|x_i\right] = P\left[\text{D}_i = 1|x_i\right] = g_0(x_i) + [g_1(x_i) - g_0(x_i)]\text{T}_i,$$

where

$$\text{T}_i = 1(x_i \geq x_0).$$

The dummy variable T_i indicates the point of discontinuity in $E\left[\text{D}_i|x_i\right]$.

Fuzzy RD leads naturally to a simple 2SLS estimation strategy. Assuming that $g_0(x_i)$ and $g_1(x_i)$ can be described by *pth*-order polynomials as in (6.1.4), we have

$$\begin{aligned} E\left[\text{D}_i|x_i\right] &= \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^2 + ... + \gamma_{0p}x_i^p \\ &\quad + [\gamma_0^* + \gamma_1^* x_i + \gamma_2^* x_i^2 + ... + \gamma_p^* x_i^p]\text{T}_i \\ &= \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^2 + ... + \gamma_{0p}x_i^p \\ &\quad + \gamma_0^*\text{T}_i + \gamma_1^* x_i\text{T}_i + \gamma_2^* x_i^2\text{T}_i + ... + \gamma_p^* x_i^p\text{T}_i. \end{aligned} \tag{6.2.1}$$

From this we see that T_i, as well as the interaction terms $\{x_i\text{T}_i, x_i^2\text{T}_i, \ldots x_i^p\text{T}_i\}$ can be used as instruments for D_i in (6.1.4).[4]

The simplest fuzzy RD estimator uses only T_i as an instrument, without the interaction terms (with the

[4] The idea of using jumps in the probability of assignment as a source of identifying information appears to originate with Trochim (1984), although the IV interpretation came later. Not everyone agrees that fuzzy RD is IV, but this view is catching on. In a recent history of the RD idea, Cook (2008) writes about the fuzzy design: "In many contexts, the cutoff value can function as an IV and engender unbiased causal conclusions . . . fuzzy assignment does not seem as serious a problem today as earlier."

interaction terms in the instrument list, we might also like to allow for interactions in the second stage as in 6.1.6). The resulting just-identified IV estimator has the virtues of transparency and good finite-sample properties. The first stage in this case is

$$\mathrm{D}_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + ... + \gamma_p x_i^p + \pi \mathrm{T}_i + \xi_{1i}, \tag{6.2.2}$$

where T_i is the excluded instrument that provides identifying power with a first-stage effect given by π.

The fuzzy RD reduced form is obtained by substituting (6.2.2) into (6.1.4):

$$\mathrm{Y}_i = \mu + \kappa_1 x_i + \kappa_2 x_i^2 + ... + \kappa_p x_i^p + \rho\pi \mathrm{T}_i + \xi_{2i}, \tag{6.2.3}$$

where $\mu = \alpha + \rho\gamma_0$ and $\kappa_j = \beta_1 + \rho\gamma_j$ for $j = 1, ..., p$. As with sharp RD, identification in the fuzzy case turns on the ability to distinguish the relation between Y_i and the discontinuous function, $\mathrm{T}_i = 1(x_i \geq x_0)$, from the effect of polynomial controls included in the first and second stage. In one of the first RD studies in applied econometrics, van der Klaauw (2002) used a fuzzy design to evaluate the effects of university financial aid awards on college enrollment. In van der Klaauw's study, D_i is the size of the financial aid award offer, and T_i is a dummy variable indicating applicants with an ability index above pre-determined award-threshold cutoffs.[5]

Fuzzy RD estimates with treatment effects that change as a function of x_i can be constructed by 2SLS estimation of an equation with treatment-covariate interactions. Here, the second stage model with interaction terms is the same as (6.1.6), while the first stage is similar to (6.2.1), except that to match the second-stage parametrization, we center polynomial terms at x_0. In this case, the excluded instruments are $\{\mathrm{T}_i, \tilde{x}_i\mathrm{T}_i, \tilde{x}_i^2\mathrm{T}_i, \ldots \tilde{x}_i^p\mathrm{T}_i\}$ while the variables $\{\mathrm{D}_i, \tilde{x}_i\mathrm{D}_i, \mathrm{D}_i\tilde{x}_i{}^2, \ldots \mathrm{D}_i\tilde{x}_i{}^p\}$ are treated as endogenous. The first stage for D_i becomes

$$\begin{aligned} \mathrm{D}_i \quad = \quad & \gamma_{00} + \gamma_{01}\tilde{x}_i + \gamma_{02}\tilde{x}_i^2 + ... + \gamma_{0p}\tilde{x}_i^p \\ & + \gamma_0^*\mathrm{T}_i + \gamma_1^*\tilde{x}_i\mathrm{T}_i + \gamma_2^*\tilde{x}_i^2\mathrm{T}_i + ... + \gamma_p^*\tilde{x}_i^p\mathrm{T}_i + \xi_{1i}. \end{aligned} \tag{6.2.4}$$

An analogous first stage is constructed for each of the polynomial interaction terms in the set $\{\tilde{x}_i\mathrm{D}_i, \mathrm{D}_i\tilde{x}_i{}^2, \ldots \mathrm{D}_i\tilde{x}_i{}^p\}$.[6]

The nonparametric version of fuzzy RD consists of IV estimation in a small neighborhood around the discontinuity. The reduced-form conditional expectation of Y_i near x_0 is

[5] van der Klaauw's original working paper circulated in 1997. Note that the fact that (6.2.2) is only an approximation of $E[\mathrm{D}_i|x_i]$ is not very important; second-stage estimates are still consistent.

[6] Alternately, center neither the first or second stage. In this case, however, ρ no longer captures the treatment effect at the cutoff.

$$E\left[\text{Y}_i|x_0 < x_i < x_0 + \delta\right] - E\left[\text{Y}_i|x_0 - \delta < x_i < x_0\right] \simeq \rho\gamma_0^*.$$

Similarly, for the first stage for D_i, we have

$$E\left[\text{D}_i|x_0 < x_i < x_0 + \delta\right] - E\left[\text{D}_i|x_0 - \delta < x_i < x_0\right] \simeq \gamma_0^*.$$

Therefore

$$\lim_{\delta \to 0} \frac{E\left[\text{Y}_i|x_0 < x_i < x_0 + \delta\right] - E\left[\text{Y}_i|x_0 - \delta < x_i < x_0\right]}{E\left[\text{D}_i|x_0 < x_i < x_0 + \delta\right] - E\left[\text{D}_i|x_0 - \delta < x_i < x_0\right]} = \rho. \tag{6.2.5}$$

The sample analog of (6.2.5) is a Wald estimator of the sort discussed in Section **??**, in this case using T_i as an instrument for D_i in a $\delta-$neighborhood of x_0. As with other dummy-variable instruments, the result is a local average treatment effect. In particular, the Wald estimand for fuzzy RD captures the causal effect on compliers defined as individuals whose treatment status changes as we move the value of x_i from just to the left of x_0 to just to the right of x_0. This interpretation of fuzzy RD was introduced by Hahn, Todd, and van der Klaauw (2001). Note, however, that there is another sense in which this version of LATE is local: the estimates are for compliers with $x_i = x_0$, a feature of sharp nonparametric estimates as well.

Finally, note that as with the nonparametric version of sharp RD, the finite-sample behavior of the sample analog of (6.2.5) is not likely to be very good. Hahn, Todd, and van der Klaauw (2001) develop a nonparametric IV procedure using local linear regression to estimate the top and bottom of the Wald estimator with less bias. This takes us back to a 2SLS model with linear or polynomial controls, but the model is fit in a discontinuity sample using a data-driven bandwidth. The idea of using discontinuity samples informally also applies in this context: start with a parametric 2SLS setup in the full sample, say, based on (6.1.4). Then restrict the sample to points near the discontinuity and get rid of most or all of the polynomial controls. Ideally, 2SLS estimates in the discontinuity samples with few controls will be broadly consistent with the more precise estimates constructed using the larger sample.

Angrist and Lavy (1999)use a fuzzy RD research design to estimate the effects of class size on children's test scores, the same question addressed by the STAR experiment discussed in Chapter 2. Fuzzy RD is an especially powerful and flexible research design, a fact highlighted by the Angrist and Lavy study, which generalizes fuzzy RD in two ways relative to the discussion above. First, the causal variable of interest, class size, takes on many values. So the first stage exploits jumps in *average* class size instead of probabilities. Second, the Angrist and Lavy (1999) research design uses multiple discontinuities.

The Angrist and Lavy study begins with the observation that class size in Israeli schools is capped at 40. Students in a grade with up to 40 students can expect to be in classes as large as 40, but grades with 41 students are split into two classes, grades with 81 students are split into three classes, and so on. Angrist

and Lavy call this "Maimonides Rule" since a maximum class size of 40 was first proposed by the medieval Talmudic scholar Maimonides. To formalize Maimonides Rule, let m_{sc} denote the predicted class size (in a given grade) assigned to class c in school s, where enrollment in the grade is denoted e_s. Assuming grade cohorts are split up into classes of equal size, the predicted class size that results from a strict application of Maimonides' Rule is

$$m_{sc} = \frac{e_s}{int[\frac{(e_s-1)}{40}] + 1}$$

where $int(x)$ is the integer part of a real number, x. This function, plotted with dotted lines in Figure 6.2.1 for fourth and fifth graders, has a sawtooth pattern with discontinuities (in this case, sharp drops in predicted class size) at integer multiples of 40. At the same time, m_{sc} is clearly an increasing function of enrollment, e_s, making the enrollment variable an important control.

Angrist and Lavy exploit the discontinuities in Maimonides Rule by constructing 2SLS estimates of an equation like

$$\mathrm{Y}_{isc} = \alpha_0 + \alpha_1 pd_s + \beta_1 e_s + \beta_2 e_s^2 + ... + \beta_p e_s^p + \rho n_{sc} + \eta_{isc}, \tag{6.2.6}$$

where Y_{isc} is $i's$ test score in school s and class c, n_{sc} is the size of this class, and e_s is enrollment. In this version of fuzzy RD, m_{sc} plays the role of T_i, e_s plays the role of x_i, and class size, n_{sc} plays the role of D_i. Angrist and Lavy also include a non-enrollment covariate, pd_s, to control for the proportion of students in the school from a disadvantaged background. This is not necessary for RD, since the only source of omitted variables bias in the RD model is e_s, but it makes the specification comparable to the model used to construct a corresponding set of OLS estimates.[7]

Figure 6.2.1 from Angrist and Lavy (1999) plots the average of actual and predicted class sizes against enrollment in fourth and fifth grade. Maimonides' Rule does not predict class size perfectly because some schools split grades at enrollments lower than 40. This is what makes the RD design fuzzy. Still, there are clear drops in class size at enrollment levels of enrollment levels of 40, 80, and 120. Note also that the m_{sc} instrument neatly combines both discontinuities and slope-discontinuity interactions such as $\tilde{x}_i \mathrm{T}_i$ in (6.2.4) in a single variable. This compact parametrization comes from a specific understanding of the institutions and rules that determine Israeli class size.

Estimates of equation (6.2.6) for fifth-grade Math scores are reported in Table 6.2.1, beginning with OLS. With no controls, there is a strong positive relationship between class size and test scores. Most of this vanishes however, when the percent disadvantaged in the school is included as a control. The correlation between class size and test scores shrinks to insignificance when enrollment is added as an additional control, as can be seen in column 3. Still, there is no evidence that smaller classes are better, as we might believe based on the results from the Tennessee STAR randomized trial.

[7] The Angrist and Lavy (1999) study differs modestly from the description here in that the data used to estimate equation (6.2.6) are class averages. But since the covariates are all defined at the class or school level, the only difference between student-level and class-level estimation is the implicit weighting by number of students in the student-level estimates.

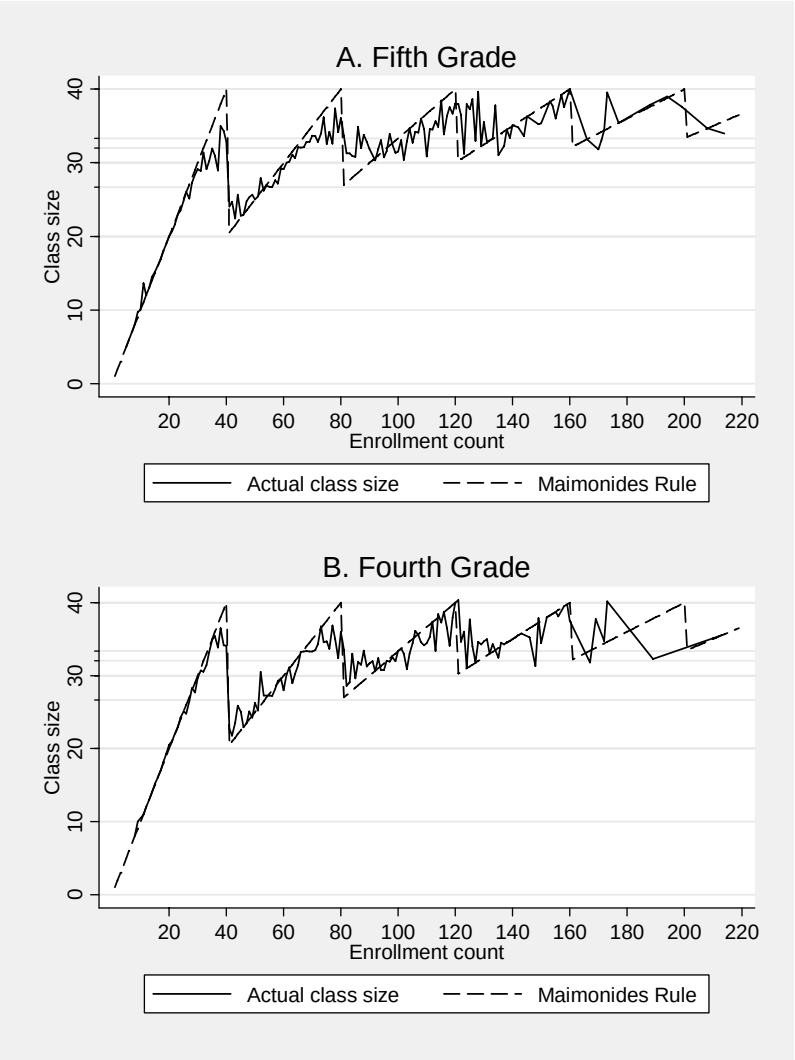

Figure 6.2.1: The fuzzy-RD first-stage for regression-discontinuity estimates of the effect of class size on pupils' test scores (from Angrist and Lavy, 1999)

In contrast with the OLS estimates in column 3, 2SLS estimates of similar specification using m_{sc} as an instrument for n_{sc} strongly suggest that smaller classes increase test scores. These results, reported in column 4 for models that include a linear enrollment control and in column 5 for models that include a quadratic enrollment control range from -.23 to -.26 with standard error around .1. These results suggest a 7-student reduction in class size (as in Tennessee STAR) raises Math scores by about 1.75 points, for an effect size of $.18\sigma$, where σ is the standard deviation of class average scores. This is not too far from the Tennessee estimates.

Importantly, the functional form of the enrollment control does not seem to matter very much (though estimates with no controls - not reported in the table - come out much smaller and insignificant). Columns 6 and 7 check the robustness of the main findings using a +/-5 discontinuity sample. Not surprisingly, these results are much less precise than those reported in columns 5 and 6 since they were estimated with only about one-quarter of the data used to construct the full-sample estimates. Still, they bounce around the -.25 mark. Finally, the last column shows the results of estimation using an even narrower discontinuity sample limited to schools with plus or minus an enrollment of 3 students around the discontinuities at 40, 80, and 120 (with dummy controls for which of these discontinuities is relevant). These are Wald estimates in the spirit of Hahn, Todd, and van der Klaauw (2001) and formula (6.2.5); the instrument used to construct these estimates is a dummy for being in a school with enrollment just to the right of the relevant discontinuity. The result is an imprecise -.270 (s.e.=.281), but still strikingly similar to the other estimates in the table. This set of estimates illustrates the high price to be paid in terms of precision when we shrink the sample around the discontinuities. Happily, however, the picture that emerges from Table (6.2.1) is fairly clear.

Table 6.2.1: OLS and fuzzy RD estimates of the effects of class size on fifth grade math scores

	OLS			2SLS				
				Full sample		Discontinuity samples		
						+/- 5		+/- 3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mean score		67.3		67.3		67.0		67.0
(s.d.)		(9.6)		(9.6)		(10.2)		(10.6)
Regressors								
Class size	.322	.076	.019	-.230	-.261	-.185	-.443	-.270
	(.039)	(.036)	(.044)	(.092)	(.113)	(.151)	(.236)	(.281)
Percent disadvantaged		-.340	-.332	-.350	-.350	-.459	-.435	
		(.018)	(.018)	(.019)	(.019)	(.049)	(.049)	
Enrollment			.017	.041	.062		.079	
			(.009)	(.012)	(.037		(.036)	
Enrollment squared/100					-.010			
					(.016)			
Segment 1								-12.6
(enrollment 36-45)								(3.80)
Segment 2								-2.89
(enrollment 76-85)								(2.41)
Root MSE	9.36	8.32	8.30	8.40	8.42	8.79	9.10	10.2
R-squared	.048	.249	.252					
N		2,018		2,018		471		302

Notes: Adapted from Angrist and Lavy (1999). The table reports estimates of equation (6.2.6) in the text using class averages. Standard errors, reported in parentheses, are corrected for within-school correlation.

Chapter 7

Quantile Regression

> Here's a prayer for you. Got a pencil? . . . 'Protect me from knowing what I don't need to know. Protect me from even knowing that there are things to know that I don't know. Protect me from knowing that I decided not to know about the things I decided not to know about. Amen.' There's another prayer that goes with it. 'Lord, lord, lord. Protect me from the consequences of the above prayer.'
>
> Douglas Adams, *Mostly Harmless* (1995)

Rightly or wrongly, 95 percent of applied econometrics is concerned with averages. If, for example, a training program raises average earnings enough to offset the costs, we are happy. The focus on averages is partly because obtaining a good estimate of the average causal effect is hard enough. And if the dependent variable is a dummy for something like employment, the mean describes the entire distribution. But many variables, like earnings and test scores, have continuous distributions. These distributions can change in ways not revealed by an examination of averages, for example, they can spread out or become more compressed. Applied economists increasingly want to know what's happening to an entire distribution, to the relative winners and losers, as well as to averages.

Policy-makers and labor economists have been especially concerned with changes in the wage distribution. We know, for example, that flat average real wages are only a small part of what's been going on in the labor market for the past 25 years. Upper earnings quantiles have been increasing, while lower quantiles have been falling. In other words, the rich are getting richer and the poor are getting poorer. But that's not all - recently, inequality has grown asymmetrically; for example, among college graduates, it's mostly the rich getting richer, with wages at the lower decile unchanging. The complete story of the changing wage distribution is fairly complicated and would seem to be hard to summarize.

Quantile regression is a powerful tool that makes the task of modeling distributions easy, even when the underlying story is complex and multi-dimensional. We can use this tool to see whether participation in a training program or membership in a labor union affects earnings inequality as well as average earnings. We

can also check for interactions, like whether and how the relation between schooling and inequality has been changing over time. Quantile regression works very much like conventional regression: confounding factors can be held fixed by including covariates; interaction terms work the same as with regular regression, too. And sometimes we can even use instrumental variables methods to estimate causal effects on quantiles when a selection-on-observables story seems implausible.

7.1 The Quantile Regression Model

The starting point for quantile regression is the conditional quantile function (CQF). Suppose we are interested in the distribution of a continuously-distributed random variable, Y_i, with a well-behaved density (no gaps or spikes). Then the CQF at quantile τ given a vector of regressors, x_i, can be defined as:

$$Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) = F_Y^{-1}(\tau|\mathrm{X}_i)$$

where $F_Y(y|\mathrm{X}_i)$ is the distribution function for Y_i conditional on X_i. When $\tau = .10$, for example, $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$ describes the lower decile of Y_i given X_i, while $\tau = .5$ gives us the conditional median.[1] By looking at changes in the CQF of earnings as a function of education, we can tell whether the dispersion in earnings goes up or down with schooling. By looking at changes in the CQF of earnings as a function of education and time, we can tell whether the relationship between schooling and inequality is changing over time.

The CQF is the conditional-quantile version of the CEF. Recall that the CEF can be derived as the solution to a mean-squared error prediction problem,

$$E\left[\mathrm{Y}_i|\mathrm{X}_i\right] = \underset{m(\mathrm{X}_i)}{\arg\min} E\left[\left(\mathrm{Y}_i - m\left(\mathrm{X}_i\right)\right)^2\right].$$

In the same spirit, the CQF solves the following minimization problem,

$$Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) = \underset{q(X)}{\arg\min}\ E\left[\rho_\tau(\mathrm{Y}_i - q(\mathrm{X}_i))\right], \tag{7.1.1}$$

where $\rho_\tau(u) = (\tau - 1(u \leq 0))u$ is called the "check function" because it looks like a check-mark when you plot it. If $\tau = .5$, this becomes least absolute deviations because $\rho_{.5}(u) = \frac{1}{2}(\operatorname{sign} u)u = \frac{1}{2}|u|$. In this case, $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$ is the conditional median since the conditional median minimizes absolute deviations. Otherwise,

[1] More generally, we can define the CQF for discrete random variables and random variables with less-than-well-behaved densities as

$$Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) = \inf\{y : F_Y(y|\mathrm{X}_i) \geq \tau\}.$$

the check function weights positive and negative terms asymmetrically:

$$\rho_\tau(u) = 1(u > 0) \cdot \tau u + 1(u \leq 0) \cdot (1 - \tau)u.$$

This asymmetric weighting generates a minimand that picks out conditional quantiles (a fact that's not immediately obvious but can be proved with a little work; see Koenker, 2005).

As a practical tool, the CQF shares the disadvantages of the CEF with continuous or high-dimensional X_i: it may be hard to estimate and summarize. We'd therefore like to boil this function down to a small set of numbers, one for each element of X_i. Quantile regression accomplishes this by substituting a linear model for $q(\mathrm{X}_i)$ in (7.1.1), producing

$$\beta_\tau \equiv \arg\min_{b \in \mathbb{R}^d} \; E\left[\rho_\tau(\mathrm{Y}_i - \mathrm{X}_i'b)\right]. \tag{7.1.2}$$

The quantile regression *estimator*, $\hat{\beta}_\tau$, is the sample analog of (7.1.2). It turns out this is a linear programming problem that is fairly easy (for computers) to solve.

Just as OLS fits a linear model to Y_i by minimizing expected squared error, quantile regression fits a linear model to Y_i using the asymmetric loss function, $\rho_\tau(\cdot)$. If $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$ is in fact linear, the quantile regression minimand will find it (just as if the CEF is linear, OLS will find it). The original quantile regression model, introduced by Koenker and Bassett (1978), was motivated by the assumption that the CQF is linear. As it turns out, however, the assumption of a linear CQF is unnecessary - quantile regression is useful whether or not we believe this.

Before turning to a more general theoretical discussion of quantile regression, we illustrate the use of this tool to study the wage distribution**.** The motivation for the use of quantile regression to look at the wage distribution comes from labor economists' interest in the question of how inequality varies conditional on covariates like education and experience (see, e.g., Buchinsky, 1994). The overall gap in earnings by schooling group (e.g., the college/high-school differential) grew considerably in the 1980s and 1990s. Less clear, however, is how the wage distribution has been changing *within* education and experience groups. Many labor economists believe that increases in so-called "within-group inequality" provide especially strong evidence of fundamental changes in the labor market, not easily accounted for by changes in institutional features like the percent of workers who belong to labor unions.

Table 7.1.1: Quantile regression coefficients for schooling in the 1970, 1980, and 2000 Censuses

		Desc. Stats.		Quantile Regression Estimates					OLS Estimates	
Census	Obs.	Mean	SD	0.1	0.25	0.5	0.75	0.9	Coeff.	Root MSE
1980	65023	6.4	0.67	.074 (.002)	.074 (.001)	.068 (.001)	.070 (.001)	.079 (.001)	.072 (.001)	0.63
1990	86785	6.46	0.06	.112 (.003)	.110 (.001)	.106 (.001)	.111 (.001)	.137 (.003)	.114 (.001)	0.64
2000	97397	6.5	0.75	.092 (.002)	.105 (.001)	.111 (.001)	.120 (.001)	.157 (.004)	.114 (.001)	0.69

Notes: Adapted from Angrist, Chernozhukov, and Fernandez-Val (2006). The tables reports quantile regression estimates of the returns to schooling, with OLS estimates shown at the right for comparison. The sample includes US-born white and black men aged 40-49. Standard errors are reported in parentheses. All models control for race and potential experience. Sampling weights were used for the 2000 Census estimates.

Table 7.1.1 reports schooling coefficients from quantile regressions estimated using the 1980, 1990, and 2000 Censuses. The models used to construct these estimates control for race and a quadratic function of potential labor market experience (defined to be $age - education - 6$). The .5 quantile coefficients - for the conditional median - are very much like the OLS coefficients at the far right-hand side of the table. For example, the OLS estimate of .072 in the 1980 census is not very different from the .5 quantile coefficient of about .068 in the same data. If the conditional-on-covariates distribution of log wages is symmetric, so that the conditional median equals the conditional mean, we should expect these two coefficients to be the same. Also noteworthy is that fact that the quantile coefficients are similar across quantiles in 1980. An additional year of schooling raises median wages by 6.8 percent, with slightly higher effects on the lower and upper quartiles of .074 and .070. Although the estimated returns to schooling increased sharply between 1980 and 1990 (up to .106 at the median, with an OLS return of .114 percent), there is a reasonably stable pattern of returns across quantiles in the 1990 Census. The largest effect is on the upper decile, a coefficient of .137, while the other quantile coefficients are around .11.

We should expect to see constant coefficients across quantiles if the effect of schooling on wages amounts to what is sometimes called a "location shift.". Here, that means that as higher schooling levels raise average earnings, other parts of the wage distribution move in tandem (i.e., within-group inequality does not change). Suppose, for example, that log wages can be described by a classical linear regression model:

$$\mathrm{Y}_i \sim N(\mathrm{X}_i'\beta, \sigma_\varepsilon^2), \tag{7.1.3}$$

where $E[\mathrm{Y}_i|\mathrm{X}_i] = \mathrm{X}_i'\beta$ and $\mathrm{Y}_i - \mathrm{X}_i'\beta \equiv \varepsilon_i$ is a Normally distributed error with constant variance σ_ε^2. Homoskedasticity means the conditional distribution of log wages is no more spread out for college graduates than for high school graduates. The implications of the linear homoskedastic model for quantiles are apparent from the fact that

$$P[\mathrm{Y}_i - \mathrm{X}_i'\beta < \sigma_\varepsilon \Phi^{-1}(\tau)|\mathrm{X}_i] = \tau,$$

where $\Phi^{-1}(\tau)$ is the inverse of the standard Normal CDF. From this we conclude that $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) = \mathrm{X}_i'\beta + \Phi^{-1}(\tau)$. In other words, apart from the changing intercept, all quantile regression coefficients are the same. The results in Table 7.1.1 for 1980 and 1990 are not too far from this stylized representation.

In contrast with the simple pattern in 1980 and 1990 Census data, quantile regression estimates from the 2000 Census differ markedly across quantiles, especially in the right tail. An additional year of schooling raises the lower decile of wages by 9.2 percent and the median by 11.1 percent. In contrast, an additional year of schooling raises the upper decile by 15.7 percent. Thus, in addition to increases in overall inequality in the 1980s and 1990s (a fact we know from simple descriptive statistics), by 2000, inequality began to increase with education as well. This development is the subject of considerable discussion among labor economists, who are particularly concerned with whether it points to fundamental or institutional changes

in the labor market (see, e.g., Autor, Katz, and Kearney (2005) and Lemieux (2008)).

Again, a parametric example helps us see where an increasing pattern of quantile regression coefficients might come from. We can generate increasing quantile regression coefficients by adding heteroskedasticity to the classical Normal regression model, (7.1.3). Suppose that

$$\mathrm{Y}_i \sim N(\mathrm{X}_i'\beta, \sigma^2(\mathrm{X}_i)),$$

where $\sigma^2(\mathrm{X}_i) = (\lambda'\mathrm{X}_i)^2$ and λ is a vector of positive coefficients such that $\lambda'\mathrm{X}_i > 0$ (perhaps proportional to β, so that the conditional variance grows with the conditional mean).[2] Then

$$P[\mathrm{Y}_i - \mathrm{X}_i'\beta < \ (\lambda'\mathrm{X}_i)\Phi^{-1}(\tau)|\mathrm{X}_i] = \tau,$$

with the implication that

$$Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) = \mathrm{X}_i'\beta + (\lambda'\mathrm{X}_i)\Phi^{-1}(\tau) = \mathrm{X}_i'[\beta + \lambda\Phi^{-1}(\tau)]. \tag{7.1.4}$$

so that quantile regression coefficients increase across quantiles.

Putting the pieces together, Table 7.1.1 neatly summarizes two stories, both related to variation in within-group inequality. First, results from the 2000 Census show inequality increasing sharply with education. The increase is asymmetric, however, and appears much more clearly in the upper tail of the wage distribution. Second, this increase is a new development. In 1980 and 1990, in contrast, schooling affected the wage distribution in a manner roughly consistent with a simple location shift.[3]

7.1.1 Censored Quantile Regression

Quantile regression allows us to look at features of the conditional distribution of Y_i when part of the distribution is hidden. Suppose you have have data of the form

$$\mathrm{Y}_{i,obs} = \mathrm{Y}_i \cdot 1[\mathrm{Y}_i < c], \tag{7.1.5}$$

[2] See Card and Lemieux (1996) for an empirical example of a regression model with this sort of heteroskedasticity. Koenker and Portnoy (1996) call this a linear location-scale model.

[3] The results in table 7.1.1 include two sets of standard errors. The first are conventional standard errors, of the sort reported by Stata's qreg command (also specifying "robust"). These presume the CQF is truly linear. The formula for these is

$$\tau(1-\tau)\{E[f_{u_\tau}(0|\mathrm{X}_i)\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{X}_i']E[f_{u_\tau}(0|\mathrm{X}_i)\mathrm{X}_i\mathrm{X}_i']^{-1},$$

where $f_{u_\tau}(0|\mathrm{X}_i)$ is the conditional density of the quantile-regression residual at zero. If the residuals are homoskedastic this simplifies to $\frac{\tau(1-\tau)}{f^2_{u_\tau}(0)}E[\mathrm{X}_i\mathrm{X}_i']^{-1}$. The second set are robust to misspecification, computed using formulas in Angrist, Chernozhukov, and Fernandez-Val (2006). In this example, the impact of nonlinearity on standard errors is minor.

where $\mathrm{Y}_{i,obs}$ is what you get to see and Y_i is the variable you would like to see. The variable $\mathrm{Y}_{i,obs}$ is *censored* - information about Y_i in $\mathrm{Y}_{i,obs}$ is limited for confidentiality reasons or because it was too difficult or time-consuming to collect more information. In the CPS, for example, high earnings are topcoded to protect respondent confidentiality. This means data above the topcode are recoded to have the topcode value. Duration data may also be censored: in a study of the effects of unemployment insurance on the duration of employment, we might follow new UI claimants for up to 40 weeks. Anyone out of work for longer has an unemployment spell length that is censored at 40. Note that limited dependent variables like hours worked or medical expenditure, discussed in Section 3.4.2, are not censored; they commonly take on the value zero by their nature, just as dummy variables like employment status do.

When dealing with censored dependent variables, quantile regression can be used to estimate the effect of covariates on conditional quantiles that are below the censoring point (assuming censoring is from above). This reflects the fact that recoding earnings above the upper decile to be equal to the upper decile has no effect on the median. So if CPS topcoding affects relatively few people (as is often true), censoring has no effect on estimates of the conditional median or even β_τ for $\tau = .75$. Likewise, if less than 10 percent of the sample is censored conditional on all values of X_i, then when estimating β_τ for τ up to .9 you can simply ignore it. Alternately, you can limit the sample to values of X_i where $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$ is below c (or above, if censoring is from the bottom with $\mathrm{Y}_{i,obs} = \mathrm{Y}_i \cdot 1[\mathrm{Y}_i > c]$).

Powell (1986) formalizes this idea with the censored quantile regression estimator. Because we may not know which conditional quantiles are below the censoring point (continuing to think of top codes), Powell proposes we work with

$$Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) = \min(c, \mathrm{X}_i'\beta_\tau^c).$$

The parameter vector β_τ^c solves

$$\beta_\tau^c \equiv \arg\min_{b\in\mathbb{R}^d} \; E\{1\left[\mathrm{X}_i'\beta_\tau^c < c\right] \cdot \rho_\tau(\mathrm{Y}_i - \mathrm{X}_i'b)\big]\}. \tag{7.1.6}$$

In other words, we solve the quantile regression minimization problem for values of X_i such that $\mathrm{X}_i'\beta_\tau^c < c$. That is, we minimize the sample analog of (7.1.6). As long is there is enough uncensored data, the resulting estimates give us the quantile regression function we would have gotten had the data not been censored (assuming the conditional quantile function is, in fact, linear). And if it turns out that the conditional quantiles you are estimating are below the censoring point, then you are back to regular quantile regression.

The sample analog of (7.1.6) is no longer a linear programming problem but Buchinsky (1994) proposes a simple iterated linear programming algorithm that appears to work well. The iterations go like this: First you estimate β_τ^c ignoring the censoring. Then find the cells with $\mathrm{X}_i'\beta_\tau^c < c$. Then estimate the quantile regression again using these cells only, and so on. This algorithm is not guaranteed to converge but it appears to do so in practice. Standard errors can be bootstrapped. Buchinsky (1994) and Chamberlain

(1994) use this approach to estimate the returns to schooling for highly-experienced workers that may have earnings above the CPS topcode. The censoring adjustment tends to increase the returns to schooling for this group.

7.1.2 The Quantile Regression Approximation Property★

The CQF of log wages given schooling is unlikely to be exactly linear, so the assumptions of the original quantile regression model fail to hold in this example. Luckily, quantile regression can also be understood as giving a MMSE linear approximation to the CQF, though in this case the MMSE problem is a little more complicated and harder to derive than for the regression-CEF theorem. For any quantile index $\tau \in (0,1)$, define the quantile regression specification error as:

$$\Delta_\tau(\mathrm{X}_i, \beta_\tau) \equiv \mathrm{X}_i'\beta_\tau - Q_\tau(\mathrm{Y}_i|\mathrm{X}_i).$$

The population quantile regression vector can be shown to minimize an expected weighted average of the squared specification error, $\Delta_\tau^2(\mathrm{X}_i, \beta)$, as shown in the following theorem from Angrist, Chernozhukov, and Fernandez-Val (2006):

Theorem 7.1.1 (Quantile Regression Approximation) *Suppose that (i) the conditional density $f_Y(y|\mathrm{X}_i)$ exists almost surely, (ii) $E[\mathrm{Y}_i]$, $E[Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)]$, and $E\|\mathrm{X}_i\|$ are finite, and (iii) β_τ uniquely solves (7.1.2). Then*

$$\beta_\tau = \arg \min_{b \in \mathbb{R}^d} E\left[w_\tau(\mathrm{X}_i, b) \cdot \Delta_\tau^2(\mathrm{X}_i, b)\right], \tag{7.1.7}$$

where

$$\begin{aligned} w_\tau(\mathrm{X}_i, b) &= \int_0^1 (1-u) \cdot f_{\epsilon(\tau)}\left(u\Delta_\tau(\mathrm{X}_i, b)|\mathrm{X}_i\right) du \\ &= \int_0^1 (1-u) \cdot f_Y\left(u \cdot \mathrm{X}_i'b + (1-u) \cdot Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)|\mathrm{X}_i\right) du \geq 0 \end{aligned}$$

and $\epsilon_i(\tau)$ is a quantile-specific residual,

$$\epsilon_i(\tau) \equiv \mathrm{Y}_i - Q_\tau(\mathrm{Y}_i|\mathrm{X}_i),$$

with conditional density $f_{\epsilon(\tau)}(e|\mathrm{X}_i)$ at $\epsilon_i(\tau) = e$. Moreover, when Y_i has a smooth conditional density, we have for β in the neighborhood of β_τ:

$$w_\tau(\mathrm{X}_i, \beta) \approx 1/2 \cdot f_Y\left(Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)|\mathrm{X}_i\right). \tag{7.1.8}$$

The quantile regression approximation theorem looks complicated but the big picture is simple. We can think of quantile regression as approximating $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$, just as OLS approximates $E[\mathrm{Y}_i|\mathrm{X}_i]$. The OLS weighting function is the histogram of X_i,which we denote $\pi(\mathrm{X}_i)$. The quantile regression weighting function, implicitly given by $w_\tau(\mathrm{X}_i, \beta_\tau) \cdot \pi(\mathrm{X}_i)$, is more elaborate than $\pi(\mathrm{X}_i)$ alone (the histogram is implicitly part of the quantile regression weighting function because the expectation in (7.1.7) is over the distribution of X_i). The term $w_\tau(\mathrm{X}_i, \beta_\tau)$ involves the quantile regression vector, β_τ, but can be rewritten with β_τ partialled out so that it is a function of X_i only (see Angrist, Chernozhukov, and Fernandez-Val, 2006, for details). In any case, the quantile regression weights are approximately proportional to the density of Y_i in the neighborhood of the CQF.

The quantile regression approximation property is illustrated in Figure 7.1.1, which plots the conditional quantile function of log wages given highest grade completed using 1980 Census data. Here we take advantage of the discreteness of schooling and large census samples to estimate the CQF non-parametrically by computing the quantile of wages for each schooling level. Panels A-C plot a nonparametric estimate of $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$ along with the linear quantile regression fit for the 0.10, 0.50, and 0.90 quantiles, where X_i includes only the schooling variable and a constant. The nonparametric cell-by-cell estimate of the CQF is plotted with circles in the figure, while the quantile regression line is solid. The figure shows how linear quantile regression approximates the CQF.

It's also interesting to compare quantile regression to a histogram-weighted fit to the CQF, similar to that provided by OLS for the CEF. The histogram-weighted approach to quantile regression was proposed by Chamberlain (1994). The Chamberlain minimum distance (MD) estimator is the sample analog of the vector $\tilde{\beta}_\tau$ obtained by solving

$$\tilde{\beta}_\tau = \arg\min_{b \in \mathbb{R}^d} \; E\left[(Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) - \mathrm{X}_i'b)^2\right] = \arg\min_{b \in \mathbb{R}^d} \; E\left[\Delta_\tau^2(\mathrm{X}_i, b)\right].$$

In other words, $\tilde{\beta}_\tau$ is the slope of the linear regression of $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$ on X_i, weighted by the histogram of X_i. In contrast with quantile regression, which requires only one pass through the data, MD relies on the ability to estimate $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$ consistently in a nonparametric first step.

Figure 1 plots MD fitted values with a dashed line. The quantile regression and MD lines are close, but they are not identical because of the weighting by $w_\tau(\mathrm{X}_i, \beta_\tau)$ in the quantile regression fit. This weighting accentuates the quality of the fit at values of X_i where Y_i is more densely distributed near the CQF. Panels D-F in Figure 7.1.1 plot the overall quantile weights, $w_\tau(\mathrm{X}_i, \beta_\tau) \cdot \pi(\mathrm{X}_i)$ against X_i. The panels also show estimates of the $w_\tau(\mathrm{X}_i, \beta_\tau)$, labeled "importance weights," and their density approximations, $1/2 \cdot f_Y\left(Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)|\mathrm{X}_i\right)$. The importance weights and the density weights are similar and fairly flat. The overall weighting function looks a lot like the schooling histogram, and therefore places the highest weight on 12 and 16 years of schooling.

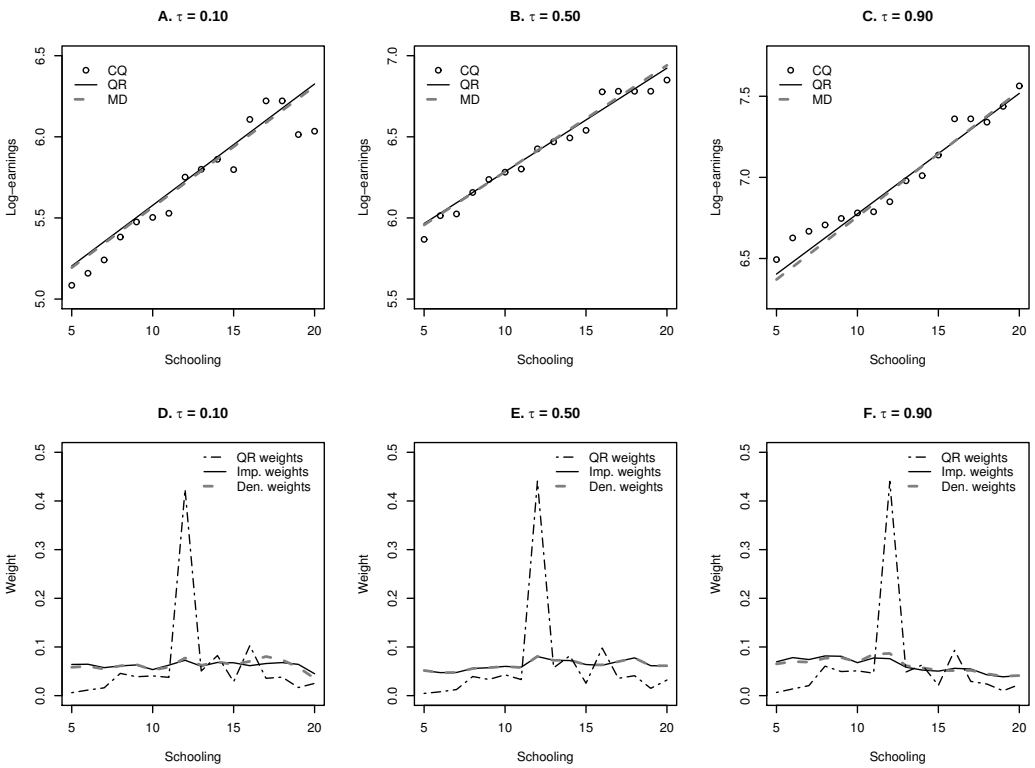

Figure 7.1.1: The quantile regression approximation property (adapted from Angrist, Chernozhukov, and Fernandez-Val, 2006). The figure shows alternative estimates of the conditional quantile function of log wages given highest grade completed using 1980 Census data, along with the implied weighting function. Panels A-C report nonparametric (CQ), quantile regression (QR) and minimum distance estimates (MD) for $\tau = .1, .5, .9$. Panels D-F show the corresponding weighting functions for QR and MD, as explained in the text.

7.1.3 Tricky Points

The language of conditional quantiles is tricky. Sometimes we talk about "quantile regression coefficients at the median," or "effects on those at the lower decile." But it's important to remember that quantile coefficients tell us about effects on *distributions* and not on *individuals*. If we discover, for example, that a training program raises the lower decile of the wage distribution, this does not necessarily mean that someone who would have been poor (i.e. at the lower decile without training) is now less poor. It only means that those who are poor in the regime with training are less poor than the poor would be in a regime without training.

The distinction between making a given set of poor people richer and changing what it means to be poor is subtle. This distinction has to do with whether we think an intervention preserves an individual's rank in the wage (or other dependent variable) distribution. If an intervention is rank-preserving, then an increase in the lower decile indeed makes those who would have been poor richer since rank preservations means relative status is unchanged. Otherwise, we can only say that the poor - defined as the group in the bottom 10 percent of the wage distribution, whoever they may be - are better off. We elaborate on this point briefly in Section 7.2, below.

A second tricky point is the transition from conditional quantiles to marginal quantiles. A link from conditional to marginal quantiles allows us to investigate the impact of changes in quantile regression coefficients on overall inequality. Suppose, for example, that quantile coefficients fan out even further with schooling, beyond what's observed in the 2000 Census. What does this imply for the ratio of upper-decile to lower-decile wages? Alternately, we can ask: how much of the overall increase in inequality (say, as measured by the ratio of upper- to lower-deciles) is explained by the fanning out of quantile regression coefficients? These sorts of questions turn out to be surprisingly difficult to answer. The difficulty has to do with the fact that all conditional quantiles are needed to pin down a particular marginal quantile (Machado and Mata, 2005). In particular, $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) = \mathrm{X}_i'\beta_\tau$ does *not* imply $Q_\tau(\mathrm{Y}_i) = Q_\tau(\mathrm{X}_i)'\beta_\tau$. This contrast this with the much more tractable expectations operator, where if $E(\mathrm{Y}_i|\mathrm{X}_i) = \mathrm{X}_i'\beta$, then by iterating expectations, we have $E(\mathrm{Y}_i) = E(\mathrm{X}_i)'\beta$.

Extracting marginal quantiles*

To show the link between conditional quantiles and marginal distributions more formally, suppose the CQF is indeed linear, so that $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) = \mathrm{X}_i'\beta_\tau$. Let $F_Y(y|\mathrm{X}_i) \equiv P[\mathrm{Y}_i < y|\mathrm{X}_i]$ with marginal distribution $F_Y(y) = P[\mathrm{Y}_i < y]$. By definition of a conditional quantile,

$$\int_0^1 1[F_Y^{-1}(\tau|\mathrm{X}_i) < y]d\tau = F_Y(y|\mathrm{X}_i). \tag{7.1.9}$$

In other words, the proportion of the population below y conditional on X_i is the same as the proportion of conditional quantiles that are below y.[4] Substituting for the CQF inside the integral,

$$F_Y(y|\mathrm{X}_i) = \int_0^1 1[\mathrm{X}_i'\beta_\tau < y]d\tau.$$

Next, we use the CDF of X_i, $F_X(x)$, to integrate and get the marginal distribution function, $F_Y(y)$:

$$F_Y(y) = \int \int_0^1 1[\mathrm{X}_i'\beta_\tau < y]d\tau dF_X(x). \tag{7.1.10}$$

Finally, marginal quantiles, say, $Q_\tau(\mathrm{Y}_i)$ for $\tau \in (0,1)$, come from inverting $F_Y(y)$:

$$Q_\tau(\mathrm{Y}_i) = \inf\{y : F_Y(y) \geq \tau\}.$$

An estimator of the marginal distribution replaces integrals with sums in (7.1.10), where the integral over quantiles comes from quantile regression estimates at, say, every .01 quantile. In a sample of size n, this is:

$$\hat{F}_Y(y) = n^{-1}\sum_i (1/100)\sum_{\tau=0}^{\tau=1} 1[\mathrm{X}_i'\hat{\beta}_\tau < y].$$

The corresponding marginal quantile estimator inverts $\hat{F}_Y(y)$.

There are a number of difficulties with this approach in practice. For one thing, you have to estimate lots of quantile regressions. Another is that the distribution theory is messy (though not insurmountable; see, e.g., Melly, 2005). Simplifying the conditional-to-marginal quantile transition is an area of active research. Gosling, Machin, and Meghir (2000) and Machado and Mata (2005) are among the first empirical studies to go from conditional to marginal quantiles. When the variable of primary interest in a quantile regression model is a dummy variable and the other regressors are seen as controls, a propensity-score type weighting scheme can be used to produce the difference in quantiles conditional on the dummy. See Firpo (2007) for the exogenous case and Frolich and Melly (2007) for a marginalization scheme that works for quantile treatment effects of the sort discussed in the next section.

7.2 Quantile Treatment Effects

The $42,000 question regarding any set of regression estimates is whether they have a causal interpretation. This is no less true for quantile regression than Ordinary Least Squares. Suppose we are interested in estimating the effect of a training program on earnings. OLS regression estimates measure the impact of the

[4] For example, if y is the conditional median, then $F_Y(y|\mathrm{X}_i) = .5$ and half of all conditional quantiles are below y. The relation (7.1.9) can be proved formally using the change of variables formula.

program on average earnings while quantile regression estimates can be used to measure the impact of the program on median earnings. In both cases, we must worry about whether the estimated program effects are contaminated by omitted variables bias.

Here too, omitted variables problems can be solved using instrumental variables, though IV methods for quantile models are a relatively new development, and are not yet as flexible as conventional 2SLS. We discuss an approach that captures the causal effect of a binary variable on quantiles (i.e., a treatment effect) using a binary instrument. The Quantile Treatment Effects (QTE) estimator, introduced in Abadie, Angrist, and Imbens (2002), relies on essentially the same assumptions as the LATE framework for average causal effects. The result is an Abadie-type weighting estimator of the causal effect of treatment on quantiles for compliers.[5]

Our discussion of the QTE estimator is based on an additive model for conditional quantiles, so that a single treatment effect is estimated. The resulting estimator simplifies to Koenker and Bassett (1978) linear quantile regression when there is no instrumenting. The relationship between QTE and quantile regression is therefore analogous to that between conventional 2SLS and OLS when the regressor of interest is a dummy.

The parameters of interest are defined as follows. For $\tau \in (0,1)$, we assume there exist $\alpha_\tau \in \mathbb{R}$ and $\beta_\tau \in \mathbb{R}^r$ such that

$$Q_\tau(\mathrm{Y}_i|\mathrm{X}_i, \mathrm{D}_i, \mathrm{D}_{1i}{>}\mathrm{D}_{0i}) = \alpha_\tau \mathrm{D}_i + \mathrm{X}_i'\beta_\tau, \tag{7.2.1}$$

where $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i,\mathrm{D}_i,\mathrm{D}_{1i}{>}\mathrm{D}_{0i})$ denotes the τ-quantile of Y_i given X_i and D_i for compliers. Thus, α_τ and β_τ are quantile regression coefficients for compliers.

Recall that D_i is independent of potential outcomes conditional on X_i and $\mathrm{D}_{1i}{>}\mathrm{D}_{0i}$, as we discussed in (4.5.2). The parameter α_τ in this model therefore gives the difference in the conditional-on-X_i quantiles of Y_{1i} and Y_{0i} for compliers. In other words,

$$Q_\tau(\mathrm{Y}_{1i}|\mathrm{X}_i, \mathrm{D}_{1i}{>}\mathrm{D}_{0i}) - Q_\tau(\mathrm{Y}_{0i}|\mathrm{X}_i, \mathrm{D}_{1i}{>}\mathrm{D}_{0i}) = \alpha_\tau \tag{7.2.2}$$

This tells us, for example, whether a training program changed the conditional median or lower decile of earnings for compliers. Note that the parameter α_τ does not tell us whether treatment changed the quantiles of the unconditional distributions of Y_{1i} and Y_{0i}. For that, we have to integrate families of quantile regression results using procedures like the one described in Section 7.1.3.

It also bears emphasizing that α_τ is not the conditional quantile of the individual treatment effects, $(\mathrm{Y}_{1i} - \mathrm{Y}_{0i})$. You might want to know, for example, whether the median treatment effect is positive. Unfortunately, questions like this are very hard to answer without making the assumptions usually invoked for causal inference.[6] Even a randomized trial with perfect compliance fails to reveal the distribution of

[5] For an alternative approach, see Chernozhukov and Hansen (2005), which allows for regressors of any type (i.e., not just dummies), but invokes a rank-invariance assumption that is unnecessary in the QTE framework.

[6] See, for example, Heckman, Smith, and Clements (1997).

($\mathrm{Y}_{1i}-\mathrm{Y}_{0i}$). This does not matter for average treatment effects since the mean of a difference is the difference in means. But all other features of the distribution of $\mathrm{Y}_{1i}-\mathrm{Y}_{0i}$ are hidden because we never get to see both Y_{1i} and Y_{0i} for any one person. The good news for applied econometricians is that the difference in marginal distributions, (7.2.2), is usually more important than the distribution of treatment effects because comparisons of aggregate economic welfare typically require only the marginal distributions of Y_{1i} and Y_{0i} and not the distribution of their difference (see, e.g., Atkinson (1970), for the traditional view). This point can be made by example without reference to quantiles. When evaluating an employment program, we are inclined to view the program favorably if it increases overall employment rates. In other words, we are happy if the average Y_{1i} is higher than the average Y_{0i}. The number of individuals who gain jobs ($\mathrm{Y}_{1i}-\mathrm{Y}_{0i}=1$) or lose jobs ($\mathrm{Y}_{1i}-\mathrm{Y}_{0i}=0$) seems like it should be of secondary interest since a good program will necessarily have more gainers than losers.

7.2.1 The QTE Estimator

The QTE estimator is motivated by the observation that, since the parameters of interest are quantile regression coefficients for compliers, they can (theoretically) be estimated consistently by running quantile regressions in the population of compliers. As always, however, the compliers population is not identifiable; we cannot list the compliers in a given data set. Nevertheless, as in Section 4.5.2, the relevant econometric minimand can be constructed using the Abadie Kappa theorem. Specifically,

$$(\alpha_\tau,\beta_\tau)=\arg\min_{a,b} E\{\rho_\tau(\mathrm{Y}_i-a\mathrm{D}_i-\mathrm{X}_i'b)|\mathrm{D}_{1i}>\mathrm{D}_{0i}\}=\arg\min_{a,b} E\{\kappa_i\rho_\tau(\mathrm{Y}_i-a\mathrm{D}_i-\mathrm{X}_i'b)\}, \tag{7.2.3}$$

where

$$\kappa_i=1-\frac{\mathrm{D}_i(1-\mathrm{Z}_i)}{1-P(\mathrm{Z}_i=1|\mathrm{X}_i)}-\frac{(1-\mathrm{D}_i)\mathrm{Z}_i}{P(\mathrm{Z}_i=1|\mathrm{X}_i)},$$

as before. The QTE estimator is the sample analog of (7.2.3).

There are a number of practical issues that arise when implementing QTE. First, κ_i must be estimated and the sampling variance induced by this first-step estimation should be reflected in the relevant asymptotic distribution theory. Abadie, Angrist, and Imbens (2002) derive the limiting distribution of the sample analog of (7.2.3) when κ_i is estimated nonparametrically. In practice, however, it is easier to bootstrap the whole procedure (i.e., beginning with the construction of estimated kappas) than to use the asymptotic formulas.

Second, κ_i is negative when $\mathrm{D}_i\neq\mathrm{Z}_i$. The kappa-weighted quantile regression minimand is therefore non-convex and no longer has a linear programming representation. This problem can be solved by working with the following minimization problem instead:

$$\min_{a,b} E\{E[\kappa_i|\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i]\rho_\tau(\mathrm{Y}_i-a\mathrm{D}_i-\mathrm{X}_i'b)\} \tag{7.2.4}$$

This minimand is derived by iterating expectations in (7.2.3). The practical difference between (7.2.3) and (7.2.4) is that the term

$$E[\kappa_i|\mathrm{Y}_i, \mathrm{D}_i, \mathrm{X}_i] = P[\mathrm{D}_{1i} > \mathrm{D}_{0i}|\mathrm{Y}_i, \mathrm{D}_i, \mathrm{X}_i]$$

is a probability and therefore between zero and one.[7] A further simplification comes from the fact that

$$E[\kappa_i|\mathrm{Y}_i, \mathrm{D}_i, \mathrm{X}_i] = 1 - \frac{\mathrm{D}_i(1 - E[\mathrm{Z}_i|\mathrm{Y}_i, \mathrm{D}_i = 1, \mathrm{X}_i)}{1 - P(\mathrm{Z}_i = 1|\mathrm{X}_i)} - \frac{(1 - \mathrm{D}_i)E[\mathrm{Z}_i|\mathrm{Y}_i, \mathrm{D}_i = 0, \mathrm{X}_i)}{P(\mathrm{Z}_i = 1|\mathrm{X}_i)}. \tag{7.2.5}$$

Angrist (2001) uses this to implement QTE via a Probit first step to estimate $E[\mathrm{Z}_i|\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i]$ separately in the $\mathrm{D}_i = 0$ and $\mathrm{D}_i = 1$ subsamples, constructing $E[\kappa_i|\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i]$ using (7.2.5), and then trimming any of the resulting estimates of $E[\kappa_i|\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i]$ that are outside the unit interval. The resulting first-step estimates of $E[\kappa_i|\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i]$ can simply be plugged in as weights when constructing quantile regression estimates in a second step using Stata's *qreg* command.[8]

Estimates of the Effect of Training on the Quantiles of Trainee Earnings

The Job Training Partnership Act was a large federal program that provided subsidized training to disadvantaged American workers in the 1980s. JTPA services were delivered at 649 sites, also called Service Delivery Areas (SDAs), located throughout the country. The original study of the labor-market impact of JTPA services was based on 15,981 people for whom continuous data on earnings (from either State unemployment insurance (UI) records or two follow-up surveys) were available for at least 30 months after random assignment.[9] There are 6,102 adult women with 30-month earnings data and 5,102 adult men with 30-month earnings data.

In our notation, Y_i is 30-month earnings, D_i indicates enrollment for JTPA services, and Z_i indicates the randomly assigned *offer* of JTPA services. A key feature of most social experiments, as with many randomized trials of new drugs and therapies, is that some participants decline the intervention being offered. In the JTPA, those offered services were not compelled to participate in training. Consequently, although the offer of subsidized training was randomly assigned, only about 60 percent of those offered training actually received JTPA services. Treatment received is therefore partly self-selected and likely to be correlated with potential outcomes. On the other hand, the randomized offer of training provides a good instrument for training received since the two are obviously correlated and the offer of treatment is independent of potential

[7] Intuitively, this is because κ_i "finds compliers". A formal statement of this result appears in Abadie, Angrist, and Imbens (2002; Lemma 3.2).

[8] Step-by-step, it goes like this:

1. Probit Z_i on Y_i and X_i separately in the $\mathrm{D}_i = 0$ and $\mathrm{D}_i = 1$ subsamples. Save these fitted values. 2. Probit Z_i on X_i in the whole sample. Save these fitted values. 3. Construct $E[\kappa_i|\mathrm{Y}_i,\mathrm{D}_i,\mathrm{X}_i]$ by plugging the two sets of fitted values into (7.2.5). Set anything less than zero to zero and anything greater than one to one. 4. Use these kappas to weight quantile regressions. 5. Bootstrap this whole procedure to construct standard errors.

[9] See Bloom *et al* (1997).

outcomes. Moreover, because of the very low percentage of individuals receiving JTPA services in the control group (less than 2 percent), effects for compliers in this case can be interpreted as effects on those who were treated (there are few always-takers).

Since training offers were randomized in the National JTPA Study, covariates (X_i) are not required to consistently estimate effects on compliers. Even in experiments like this, however, it's customary to control for covariates to correct for chance associations between treatment status and applicant characteristics and to increase precision (see Chapter 2). The covariates used here are baseline measures from the JTPA intake process. They include dummies for black and Hispanic applicants, a dummy for high-school graduates (including GED holders), dummies for married applicants, 5 age-group dummies, and a dummy for whether the applicant worked at least 12 weeks in the year preceding random assignment. Also included are dummies for the original recommended service strategy (classroom, on-the-job training (OJT), job search assistance (JSA), other) and a dummy for whether earnings data are from the second follow-up survey. Since these covariates mostly summarize coarse demographics, we can think of the quantile analysis as telling us how the JTPA experiment affected the earnings distribution within demographic groups.

As a benchmark, OLS and conventional instrumental variables (2SLS) estimates of the impact of training are reported in the first column of Table 7.2.1. The OLS training coefficient is a precisely estimated $3,754. This is the coefficient on D_i in a regression of Y_i on D_i and X_i. These estimates ignore the fact that trainees are self-selected. The 2SLS estimates in Table 7.2.1 use the randomized offer of treatment Z_i as an instrument for D_i. The 2SLS estimate for men is $1,593 with a standard error of $895, less than half the size of the corresponding OLS estimate.

Quantile regression estimates show that the gap in quantiles by trainee status is much larger (in proportionate terms) below the median than above it. This can be seen in the right-hand columns of Table 7.2.1, which reports quantile regression estimates for the .15, .25, .5, .75, and .85 quantiles. Specifically, the .85 quantile of trainee earnings is about 13 percent higher than the corresponding quantile for non-trainees, while the .15 quantile is 136 percent higher. Like the OLS estimates in the table, these quantile regression coefficients do not necessarily have a causal interpretation. Rather they provide a descriptive comparison of the earnings distributions of trainees and non-trainees.

QTE estimates of the effect of training on median earnings are similar in magnitude though less precise than the benchmark 2SLS estimates. On the other hand, the QTE estimates show a pattern very different from the quantile regression estimates, with no evidence of an impact on the .15 or .25 quantile. The estimates at low quantiles are substantially smaller than the corresponding quantile regression estimates, and they are small in absolute terms. For example, the QTE estimate (standard error) of the effect on the .15 quantile is $121 (475), while the corresponding quantile regression estimate is $1,187 (205). Similarly, the QTE estimate (standard error) of the effect on the .25 quantile for men is $702 (670), while the corresponding quantile regression estimate is $2,510 (356). Unlike the results at low quantiles, however, the QTE estimates

of effects on male earnings above the median are large and statistically significant (though still smaller than the corresponding quantile regression estimates).

The result that JTPA training for adult men did not raise the lower quantiles of their earnings is the most interesting finding arising from this analysis. This suggests that the quantile regression estimates in the top half of Table 7.2.1 are contaminated by positive selection bias. One response to this finding might be that few JTPA applicants were very well off, so that distributional effects within applicants are of less concern than the fact that the program helped many applicants overall. However, the upper quantiles of earnings were reasonably high for adults who participated in the National JTPA Study. Increasing earnings in this upper tail is therefore unlikely to have been a high priority.

Table 7.2.1: Quantile regression estimates and quantile treatment effects from the JTPA experiment

A. OLS and Quantile Regression Estimates

	OLS	Quantile				
		0.15	0.25	0.50	0.75	0.85
Training	3,754	1,187	2,510	4,420	4,678	4,806
	(536)	(205)	(356)	(651)	(937)	(1,055)
% Impact of Training	21.20	135.56	75.20	34.50	17.24	13.43
High school or GED	4,015	339	1,280	3,665	6,045	6,224
	(571)	(186)	(305)	(618)	(1,029)	(1,170)
Black	-2,354	-134	-500	-2,084	-3,576	-3,609
	(626)	(194)	(324)	(684)	(1087)	(1,331)
Hispanic	251	91	278	925	-877	-85
	(883)	(315)	(512)	(1,066)	(1,769)	(2,047)
Married	6,546	587	1,964	7,113	10,073	11,062
	(629)	(222)	(427)	(839)	(1,046)	(1,093)
Worked less than 13	-6,582	-1,090	-3,097	-7,610	-9,834	-9,951
weeks in past year	(566)	(190)	(339)	(665)	(1,000)	(1,099)
Constant	9,811	-216	365	6,110	14,874	21,527
	(1,541)	(468)	(765)	(1,403)	(2,134)	(3,896)

B. 2SLS and QTE Estimates

	2SLS	Quantile				
		0.15	0.25	0.50	0.75	0.85
Training	1,593	121	702	1,544	3,131	3,378
	(895)	(475)	(670)	(1,073)	(1,376)	(1,811)
% Impact of Training	8.55	5.19	11.99	9.64	10.69	9.02
High school or GED	4,075	714	1,752	4,024	5,392	5,954
	(573)	(429)	(644)	(940)	(1,441)	(1,783)
Black	-2,349	-171	-377	-2,656	-4,182	-3,523
	(625)	(439)	(626)	(1,136)	(1,587)	(1,867)
Hispanic	335	328	1,476	1,499	379	1,023
	(888)	(757)	(1,128)	(1,390)	(2,294)	(2,427)
Married	6,647	1,564	3,190	7,683	9,509	10,185
	(627)	(596)	(865)	(1,202)	(1,430)	(1,525)
Worked less than 13	-6,575	-1,932	-4,195	-7,009	-9,289	-9,078
weeks in past year	(567)	(442)	(664)	(1,040)	(1,420)	(1,596)
Constant	10,641	-134	1,049	7,689	14,901	22,412
	(1,569)	(1,116)	(1,655)	(2,361)	(3,292)	(7,655)

Notes: The table reports OLS, quantile regression, 2SLS, and QTE estimates of the effect of training on earnings (adapted from Abadie, Angrist, and Imbens (2002)). Assignment status is used as an instrument for training status in Panel B. All models include as covariates dummies for service strategy recommended and age group, and a dummy indicating data from a second follow-up survey. Robust standard errors are reported in parenthesis.

Chapter 8

Nonstandard Standard Error Issues

> We have normality. I repeat, we have normality.
> Anything you still can't cope with is therefore your own problem.
> Douglas Adams, *The Hitchhiker's Guide to the Galaxy* (1979)

Today, software packages routinely compute asymptotic standard errors derived under weak assumptions about the sampling process or underlying model. For example, you get regression standard errors based on formula (3.1.7) using the Stata option "robust". Robust standard errors improve on old-fashioned standard errors because the resulting inferences are asymptotically valid when the regression residuals are heteroskedastic, as they almost certainly are when regression approximates a nonlinear CEF. In contrast, old-fashioned standard errors are derived assuming homoskedasticity. The hang-up here is that robust standard errors can be misleading when the asymptotic approximation is not very good. The first part of this chapter looks at the failure of asymptotic inference with robust standard errors and some simple palliatives.

A pillar of traditional cross-section inference - and the discussion in Section 3.1.3 - is the assumption that the data are independent. Each observation is treated as a random draw from the same population, uncorrelated with the observation before or after. We understand today that this sampling model is unrealistic and potentially even foolhardy. Much as in the time-series studies common in macroeconomics, cross-section analysts must worry about correlation between observations. The most important form of dependence arises in data with a group structure - for example, the test scores of children observed within classes or schools. Children in the same school or class tend to have test scores that are correlated since they are subject to some of the same environmental and family-background influences. We call this correlation the *clustering problem*, or the *Moulton problem*, after Moulton (1986), who made it famous. A closely-related problem is correlation over time in the data sets commonly used to implement differences-in-differences estimation strategies. For example, studies of state-level minimum wages must confront the fact that state average employment rates are correlated over time. We call this the *serial correlation problem*, closely

related but distinct from the Moulton problem.

Researchers plagued by clustering and serial correlation also have to confront the fact that the simplest fix-ups for these problems, like Stata's "cluster" option, may not be very good. The asymptotic approximation relevant for clustered or serially correlated data relies on a large number of clusters or time series observations. Alas, we are rarely blessed with many clusters or long time series. The resulting inference problems are not always insurmountable, though often the best solution is to get more data. Econometric fix-ups for clustering and serial correlation are discussed in the second part of this chapter. Some of the material in this chapter is hard to work through without matrix algebra, so we take the plunge and switch to a mostly-matrix motif.

8.1 The Bias of Robust Standard Errors★

In matrix notation

$$\hat{\beta} = \left[\sum_i \mathrm{X}_i\mathrm{X}_i'\right]^{-1} \sum_i \mathrm{X}_i\mathrm{Y}_i = (X'X)^{-1}X'y,$$

where X is the $N\times\mathrm{K}$ matrix with rows X_i' and y is the $N\times 1$ vector of Y_i's. We saw in Section 3.1.3 that $\hat{\beta}$ has an asymptotically Normal distribution. We can write:

$$\sqrt{N}(\hat{\beta}-\beta) \sim N(0,\Omega)$$

where Ω is the asymptotic covariance matrix. Repeating (3.1.7), the formula for Ω in this case is

$$\Omega_r = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{X}_i'e_i^2]E[\mathrm{X}_i\mathrm{X}_i']^{-1}, \tag{8.1.1}$$

where $e_i = \mathrm{Y}_i - \mathrm{X}_i'\beta$. When residuals are homoskedastic, Ω simplifies to $\Omega_c = \sigma^2 E[\mathrm{X}_i\mathrm{X}_i']^{-1}$ where $\sigma^2 = E[e_i^2]$.

We are concerned here with the bias of robust standard errors in independent samples (i.e., no clustering or serial correlation). To simplify the derivation of bias, we assume that the regressor vector can be treated as fixed in repeated samples, as it would be if we sampled stratifying on X_i. Non-stochastic-regressors gives a benchmark sampling model that is often used to look at finite-sample distributions. It turns out that we miss little by making this assumption, while simplifying the derivations considerably.

With fixed regressors, we have

$$\Omega_r = \left(\frac{X'X}{N}\right)^{-1}\left(\frac{X'\Psi X}{N}\right)\left(\frac{X'X}{N}\right)^{-1} \tag{8.1.2}$$

where

$$\Psi = E[ee'] = diag(\psi_i)$$

is the variance matrix of residuals. Under homoskedasticity, $\psi_i = \sigma^2$ for all i and we get

$$\Omega_c = \sigma^2 \left(\frac{X'X}{N} \right)^{-1}.$$

Asymptotic standard errors are given by the square root of the diagonal elements of Ω_r and Ω_c, after removing the asymptotic normalization by dividing by N.

In practice, the asymptotic covariance matrix must be estimated. The old-fashioned or conventional variance matrix estimator is

$$\hat{\Omega}_c = (X'X)^{-1}\hat{\sigma}^2 = (X'X)^{-1} \left(\sum \frac{\widehat{e}_i^2}{N} \right),$$

where $\widehat{e}_i = \mathrm{Y}_i - \mathrm{X}_i'\hat{\beta}$ is the estimated regression residual, and

$$\hat{\sigma}^2 = \sum \frac{\widehat{e}_i^2}{N}$$

estimates the residual variance. The corresponding robust variance matrix estimator is

$$\hat{\Omega}_r = (X'X)^{-1} \left(\sum \frac{\mathrm{X}_i \mathrm{X}_i' \widehat{e}_i^2}{N} \right) (X'X)^{-1}. \tag{8.1.3}$$

We can think of the middle term as an estimator of the form $\sum \frac{\mathrm{X}_i \mathrm{X}_i' \widehat{\psi}_i}{N}$, where $\widehat{\psi}_i = \widehat{e}_i^2$ estimates ψ_i.

By the law of large numbers and Slutsky theorems, $N\hat{\Omega}_c$ converges in probability to Ω_c while $N\hat{\Omega}_r$ converges to Ω_r. But in finite samples, both variance estimators are biased. The bias in $\hat{\Omega}_c$ is well-known from classical least-squares theory and easy to correct. Less appreciated is the fact that if the residuals are homoskedastic, the robust estimator is more biased than the conventional, perhaps a lot more. From this we conclude that robust standard errors can be more misleading than conventional standard errors in situations where heteroskedasticity is modest. We also propose a rule-of-thumb that uses the maximum of old-fashioned and robust standard errors to avoid gross misjudgments of precision.

With non-stochastic regressors, we have

$$E[\hat{\Omega}_c] = (X'X)^{-1}\hat{\sigma}^2 = (X'X)^{-1} \left(\sum \frac{E(\widehat{e}_i^2)}{N} \right).$$

To analyze $E[\hat{e}_i^2]$, start by expanding $\widehat{e} = y - X\widehat{\beta}$:

$$\widehat{e} = y - X(X'X)^{-1}X'y = \left[I - X(X'X)^{-1}X' \right] (X\beta + e) = Me$$

where e is the vector of population residuals, $M = I_N - X(X'X)^{-1}X'$ is a non-stochastic residual-maker

matrix with i^{th} row m_i', and I_N is the $N \times N$ identity matrix. Then $\widehat{e}_i = m_i' e$, and

$$\begin{aligned} E\left(\widehat{e}_i^2\right) &= E\left(m_i' e e' m_i\right) \\ &= m_i' \Psi m_i \end{aligned}$$

To simplify further, write $m_i = \ell_i - h_i$ where ℓ_i is the i^{th} column of I_N and $h_i = X(X'X)^{-1}\mathrm{X}_i$, the i^{th} column of the projection matrix $H = X(X'X)^{-1}X'$. Then

$$\begin{aligned} E\left(\widehat{e}_i^2\right) &= \left(\ell_i - h_i\right)' \Psi \left(\ell_i - h_i\right) \\ &= \psi_i - 2\psi_i h_{ii} + h_i' \Psi h_i \end{aligned} \tag{8.1.4}$$

where h_{ii}, the i^{th} diagonal element of H, satisfies

$$h_{ii} = h_i' h_i = \mathrm{X}_i'(X'X)^{-1}\mathrm{X}_i. \tag{8.1.5}$$

Parenthetically, h_{ii} is called the *leverage* of the i^{th} observation. Leverage tells us how much pull a particular value of X_i exerts on the regression line. Note that the i^{th} fitted value (i^{th} element of Hy) is

$$\hat{\mathrm{Y}}_i = h_i' y = h_{ii}\mathrm{Y}_i + \sum_{j \neq i} h_{ij}\mathrm{Y}_j. \tag{8.1.6}$$

A large h_{ii} means that the i^{th} observation has a large impact on the i^{th} predicted value. In a bivariate regression with a single regressor, x_i,

$$h_{ii} = \frac{1}{N} + \frac{(x_i - \overline{x})^2}{\sum (x_j - \overline{x})^2}. \tag{8.1.7}$$

This shows that leverage increases when x_i is far the mean. In addition to (8.1.6), we know that h_{ii} is a number that lies in the interval $[0, 1]$ and that $\sum_{j=1}^{N} h_{ij} = \mathrm{K}$, the number of regressors (see, e.g., Hoaglin and Welch, 1978).[1]

Suppose residuals are homoskedastic, so that $\psi_i = \sigma^2$. Then (8.1.4) simplifies to

$$E\left(\widehat{e}_i^2\right) = \sigma^2[1 - 2h_{ii} + h_i' h_i] = \sigma^2\left(1 - h_{ii}\right) < \sigma^2.$$

[1] The property $\sum_{j=1}^{N} h_{ij} = \mathrm{K}$ comes from the fact that H is idempotent. You can also use (8.1.7) to verify that in a bivariate regression, $\sum_{j=1}^{N} h_{ij} = 2$.

So $\hat{\Omega}_c$ tends to be too small . Using the properties of h_{ii}, we can go one step further:

$$\sum \frac{E(\hat{e}_i^2)}{N} = \sigma^2 \sum \frac{1-h_{ii}}{N} = \sigma^2 \left(\frac{N-\mathrm{K}}{N}\right).$$

Thus, the bias in $\hat{\Omega}_c$ can be fixed by a simple degrees-of-freedom correction: divide by $N-\mathrm{K}$ instead of N in the formula for $\hat{\sigma}^2$, the default in most empirical variance computations.

We now want to show that under homoskedasticity the bias in $\hat{\Omega}_r$ is likely to be worse than the bias in $\hat{\Omega}_c$. The bias in the robust covariance matrix estimator is

$$E[\hat{\Omega}_r] = N(X'X)^{-1}\left(\sum \frac{\mathrm{X}_i\mathrm{X}_i' E(\hat{e}_i^2)}{N}\right)(X'X)^{-1}, \tag{8.1.8}$$

where $E\left(\hat{e}_i^2\right)$ is given by (8.1.4). Under homoskedasticity, $\psi_i = \sigma^2$ and we have $E\left(\hat{e}_i^2\right) = \sigma^2\left(1-h_{ii}\right)$ as in $\hat{\Omega}_c$. It's clear, therefore, that the bias in $\hat{e}_i^2$ tends to pull robust standard errors down. The general expression, (8.1.8), is hard to evaluate, however. Chesher and Jewitt (1987) show that as long as there is not "too much" heteroskedasticity, robust standard errors based on $\hat{\Omega}_r$ are indeed biased downwards.[2]

How do we know that $\hat{\Omega}_r$ is likely to be more biased than $\hat{\Omega}_c$? Partly this comes from Monte Carlo evidence (e.g., MacKinnon and White, 1985, and our own small study, discussed below). We also prove this for a bivariate example, where the single regressor, $\tilde{x}_i$, is assumed to be in deviations-from-means form, so there is a single coefficient. In this case, the estimator of interest is $\hat{\beta}_1 = \frac{\sum \tilde{x}_i \mathrm{Y}_i}{\sum \tilde{x}_i^2}$ and the leverage is $h_{ii} = \frac{\tilde{x}_i^2}{\sum \tilde{x}_i^2}$ (we lose the $\frac{1}{N}$ term in (8.1.7) by partialling out the constant). Let $s_x^2 = \frac{\sum \tilde{x}_i^2}{N}$. For the conventional covariance estimator, we have

$$E[\hat{\Omega}_c] = \frac{\sigma^2}{N s_x^2}\left[\frac{\sum(1-h_{ii})}{N}\right] = \frac{\sigma^2}{N s_x^2}\left[1-\frac{1}{N}\right],$$

so the bias here is small. A simple calculation using (8.1.8) shows that under heteroskedasticity, the robust estimator has expectation:

$$E[\hat{\Omega}_r] = \frac{\sigma^2}{N s_x^2}\sum \frac{(1-h_{ii})}{N}\left(\frac{\tilde{x}_i^2}{s_x^2}\right) = \frac{\sigma^2}{N s_x^2}\sum (1-h_{ii})h_{ii} = \frac{\sigma^2}{N s_x^2}\left[1-\sum h_{ii}^2\right].$$

The bias of $\hat{\Omega}_r$ is therefore worse than the bias of $\hat{\Omega}_c$ if $\sum h_{ii}^2 > \frac{1}{N}$, as it is by Jensen's inequality unless the regressor has constant leverage, in which case $h_{ii}=\frac{1}{N}$ for all i.[3]

We can reduce the bias in $\hat{\Omega}_r$ by trying to get a better estimator of ψ_i, say $\widehat{\psi}_i$. The estimator $\hat{\Omega}_r$ sets $\widehat{\psi}_i = \hat{e}_i^2$, the estimator proposed by White (1980a) and our starting point in this section. Here is a summary

[2] In particular, as long as the ratio of the largest ψ_i to the smallest ψ_i is less than 2, robust standard errors are biased downwards.

[3] Think of h_{ii} as a random variable with a uniform distribution in the sample. Then

$$E[h_{ii}] = \frac{\sum h_{ii}}{N} = \frac{1}{N}$$

of the proposals explored in MacKinnon and White (1985):

$$\begin{aligned} HC_0 &: \quad \widehat{\psi}_i = \widehat{e}_i^2 \\ HC_1 &: \quad \widehat{\psi}_i = \frac{N}{N-\mathrm{K}}\widehat{e}_i^2 \\ HC_2 &: \quad \widehat{\psi}_i = \frac{1}{1-h_{ii}}\widehat{e}_i^2 \\ HC_3 &: \quad \widehat{\psi}_i = \frac{1}{(1-h_{ii})^2}\widehat{e}_i^2. \end{aligned}$$

HC_1 is a simple degrees of freedom correction as is used for $\hat{\Omega}_c$. HC_2 uses the leverage to give an unbiased estimate of the variance estimate of the i^{th} residual when the residuals are homoskedastic, while HC_3 approximates a jackknife estimator.[4] In the applications we've seen, the estimated standard errors tend to get larger as we go down the list, but this is not a theorem.

Time out for the Bootstrap

Bootstrapping is a resampling scheme that offers an alternative to inference based on asymptotic formulas. A bootstrap sample is a sample drawn from our own data. In other words, if we have a sample of size N, we treat this sample as if it were the population and draw repeatedly from it (with replacement). The bootstrap standard error is the standard deviation of an estimator across many draws of this sort. Intuitively, we expect the sampling distribution constructed by sampling from our own data to provide a good approximation to the sampling distribution we are after.

There are many ways to bootstrap regression estimates. The simplest is to draw pairs of $\{\mathrm{Y}_i, \mathrm{X}_i\}$-values, sometimes called the "pairs bootstrap" or a "nonparametric bootstrap". Alternatively, we can keep the X_i-values fixed, draw from the distribution of residuals $(\widehat{e}_i)$, and create a new estimate of the dependent variable based on the predicted value and the residual draw for the particular observation. This procedure, which is a type of "parametric bootstrap", mimics a sample drawn with non-stochastic regressors and ensures that X_i and the regression residuals are independent. On the other hand, we don't want independence if we're interested in standard errors under heteroskedasticity. An alternative residual bootstrap, called the "wild bootstrap", draws $\mathrm{X}_i'\hat{\beta} + \widehat{e}_i$ (which, of course, is just the original Y_i) with probability 0.5, and $\mathrm{X}_i'\hat{\beta} - \widehat{e}_i$ otherwise (see, e.g., Mammen 1993 and Horowitz, 1997). This preserves the relationship between residual variances and X_i observed in the original sample.

---

and

$$E[h_{ii}^2] = \frac{\sum h_{ii}^2}{N} > (E[h_{ii}])^2 = \left(\frac{1}{N}\right)^2$$

by Jensen's inequality unless h_{ii} is constant. Therefore $\sum h_{ii}^2 > \frac{1}{N}$.The constant-leverage case occurs when $\tilde{x}_i = +/-\kappa$, for some constant, κ.

[4] A jackknife variance estimator estimates sampling variance from the empirical distribution generated by omitting one observation at a time. Stata computes HC_1, HC_2, and HC_3. You can also use a trick suggested by Messer and White (1984): divide Y_i and X_i by $\sqrt{\widehat{\phi}_i}$ and instrument the transformed model by $X_i/\sqrt{\widehat{\phi}_i}$ for your preferred choice of $\widehat{\phi}_i$.

Bootstrapping is useful for two reasons. First, in some cases the asymptotic distribution of an estimator can be hard to compute (e.g., the asymptotic distributions of quantile regression estimates involve unknown densities). Bootstrapping provides a computer-intensive but otherwise straightforward computational strategy. Not all asymptotic distributions are approximated by the bootstrap, but it seems to work well for the simple estimators we care about. Second, under some circumstances, the sampling distribution obtained via bootstrap may be closer to the finite-sample distribution of interest than the asymptotic approximation - statisticians call this property asymptotic refinement.

Here, we are mostly interested in the bootstrap because of asymptotic refinement. The asymptotic distribution of regression estimates is easy enough to compute, but we worry that the estimators HC_0 - HC_3 are biased. As a rule, bootstrapping provides an asymptotic refinement when applied to test statistics that have asymptotic distributions which do not depend on any unknown parameters (see, e.g., Horowitz, 2001). Such test statistics are said to be asymptotically pivotal. An example is a t-statistic: this is asymptotically standard normal. Regression coefficients are not asymptotically pivotal; they have an asymptotic distribution which depends on the unknown residual variance.

The upshot is that if you want better finite-sample inference for regression coefficients, you should bootstrap t-statistics. That is, you calculate the t-statistic in each bootstrap sample and compare the analogous t-statistic from your original sample to this bootstrap "t"-distribution. A hypothesis is rejected if the absolute value of the original t-statistic is above, say, the 95^{th} percentile of the absolute values from the bootstrap distribution.

Theoretical appeal notwithstanding, as applied researchers, we don't like the idea of bootstrapping pivotal statics very much. This is partly because we're not only (or even primarily) interested in formal hypothesis testing: we like to see the standard errors in parentheses under our regression coefficients. These provide a summary measure of precision that can be used to construct confidence intervals, compare estimators, and test any hypothesis that strikes us, now or later. We can certainly calculate standard errors from bootstrap samples but this promises no asymptotic refinement. In our view, therefore, practitioners worried about the finite-sample behavior of robust standard errors should focus on bias corrections like HC_1-HC_3. We especially like the idea of taking the larger of the conventional standard error (with degrees of freedom correction) and one of these three.

An Example

For further insight into the differences between robust covariance estimators, we analyze a simple but important example that has featured in the earlier chapters in this book. Suppose you are interested in an estimate of β_1 in the model

$$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{D}_i + \varepsilon_i, \tag{8.1.9}$$

where D_i is a dummy variable. The OLS estimate of β_1 is the difference in the means between those with D_i switched on and off. Denoting these subsamples by the subscripts 1 and 0, we have

$$\widehat{\beta}_1 = \overline{\mathrm{Y}}_1 - \overline{\mathrm{Y}}_0.$$

For the purposes of this derivation we think of D_i as non-random, so that $\sum \mathrm{D}_i = N_1$ and $\sum(1-\mathrm{D}_i) = N_0$ are fixed. Let $r = N_1/N$.

We know something about the finite-sample behavior of $\widehat{\beta}_1$from statistical theory. If Y_i is Normal with equal but unknown variance in both the $\mathrm{D}_i = 1$ and $\mathrm{D}_i = 0$ populations, then the conventional t–statistic for $\widehat{\beta}_1$ has a t-distribution. This is the classic two sample t-test. Heteroskedasticity in this context means that the variances in the $\mathrm{D}_i = 1$ and $\mathrm{D}_i = 0$ populations are different. In this case, the testing problem in small samples becomes surprisingly intractable: the exact small sample distribution for even this simple problem is unknown.[5] The robust covariance estimators HC_0 - HC_3 give asymptotic approximations to the unknown finite-sample distribution for the case of unequal variances.

The differences between HC_0 - HC_3 are differences in how the sample variances in the two groups defined by D_i are processed. Define $S_j^2 = \sum_{\mathrm{D}_i=j} (\mathrm{Y}_i - \overline{\mathrm{Y}}_j)^2$ for $j = 0, 1$. The leverage in this example is

$$h_{ii} = \begin{array}{ll} 1/N_0 & \text{if } \mathrm{D}_i = 0 \\ 1/N_1 & \text{if } \mathrm{D}_i = 1 \end{array}.$$

Using this, it's straightforward to show that the five variance estimators we've been discussing are

$$\begin{aligned}
\textit{Conventional} &: \frac{N}{N_0 N_1}\left(\frac{S_0^2 + S_1^2}{N-2}\right) = \frac{1}{Nr(1-r)}\left(\frac{S_0^2 + S_1^2}{N-2}\right) \\
HC_0 \text{ (White, 1980)} &: \frac{S_0^2}{N_0^2} + \frac{S_1^2}{N_1^2} \\
HC_1 &: \frac{N}{N-2}\left(\frac{S_0^2}{N_0^2} + \frac{S_1^2}{N_1^2}\right) \\
HC_2 &: \frac{S_0^2}{N_0\left(N_0-1\right)} + \frac{S_1^2}{N_1\left(N_1-1\right)} \\
HC_3 &: \frac{S_0^2}{\left(N_0-1\right)^2} + \frac{S_1^2}{\left(N_1-1\right)^2}.
\end{aligned}$$

The conventional estimator pools subsamples: this is efficient when the two variances are the same. The White (1980a) estimator, HC_0, adds separate estimates of the sampling variances of the means, using the consistent (but biased) variance estimators, $\frac{S_j^2}{N_j}$. The HC_2 estimator uses unbiased estimators of the sample sample variance for each group, since it makes the correct degrees of freedom correction. HC_1 makes a degrees of freedom correction outside the sum, which will help but is generally not quite correct. Since we know HC_2 to be the unbiased estimate of the sampling variance under homoskedasticity, HC_3 must be too

[5] This is known as the Behrens-Fisher problem (see e.g. DeGroot and Schervish, 2001, ch. 8).

big. Note that with $r = 0.5$, a case where the regression design is said to be balanced, the conventional estimator equals HC_1 and all five estimators differ little.

A small Monte Carlo study based on (8.1.9) illustrates the pluses and minuses of the estimators and the extent to which a simple rule of thumb goes a long way towards ameliorating the bias of the HC class. We choose $N = 30$ to highlight small sample issues, and $r = 0.9$, which implies $h_{ii} = 10/N = 1/3$ if $\text{D}_i = 1$. This is a highly unbalanced design. We draw

$$\varepsilon_i \sim \begin{cases} N(0, \sigma^2) & \text{if } \text{D}_i = 0 \\ N(0, 1) & \text{if } \text{D}_i = 1 \end{cases}$$

and report results for three cases. The first has lots of heteroskedasticity with $\sigma = 0.5$, while the second has relatively little heteroskedasticity, with $\sigma = 0.85$. No heteroskedasticity is the benchmark case.

Table 8.1.1 displays the results. Columns (1) and (2) report means and standard deviations of the various standard error estimators across 25,000 replications of the sampling experiment. The standard deviation of $\widehat{\beta_1}$ is the sampling variance we are trying to measure. With lots of heteroskedasticity, as in the upper panel of the table, Conventional standard errors are badly biased and, on average, only about half the size of the Monte Carlo sampling variance that constitutes our target. On the other hand, while the robust standard errors perform better, except for HC_3, they are still too small.[6]

The standard errors are themselves estimates and have considerable sampling variability. Especially noteworthy is the fact that the robust standard errors have much higher sampling variability than the OLS standard errors, as can be seen in column 2.[7] The sampling variability further increases when we attempt to reduce bias by dividing the residuals by $1 - h_{ii}$ or $(1 - h_{ii})^2$. The worst case is HC_3, with a standard deviation about 50% above that of the White (1980a) standard error, HC_0.

The last two columns in the table show empirical rejection rates in a nominal 5% test for the hypothesis $\widehat{\beta}_1 = \beta^*$, where β^* is the population parameter (equal to zero, in this case). The test statistics are compared with a Normal distribution and to a t-distribution with $N-2$ degrees of freedom. Rejection rates are far too high for all tests, even HC_3. Using a t-distribution rather than a Normal distribution helps only marginally.

The results with little heteroskedasticity, reported in the second panel, show that conventional standard errors are still too low; this bias is now in the order of 15%. HC_0 and HC_1 are also too small, about like before in absolute terms, though they now look worse relative to the conventional standard errors. The HC_2 and HC_3 standard errors are still larger than the conventional standard errors, on average, but

[6] Notice that HC_2 is an unbiased estimator of the *sampling variance,* while the mean of the HC_2 *standard errors* across sampling experiments (0.52) is still below the standard deviation of $\widehat{\beta}$ (0.59). This comes from the fact that the standard error is the square root of the sampling variance, the sampling variance is itself estimated and hence has sampling variability, and the square root is a concave function.

[7] The large sampling variance of robust standard error estimators is noted by Chesher and Austin (1991). Kauermann and Carroll (2001) propose an adjustment to confidence intervals to correct for this.

empirical rejection rates are higher for these two than for conventional standard errors. This means the robust standard errors are sometimes too small "by accident," an event that happens often enough to inflate rejection rates so that they exceed the conventional rejection rates.

The lesson we can take a away from this is that robust standard errors are no panacea. They can be smaller than conventional standard errors for two reasons: the small sample bias we have discussed and the higher sampling variance of these standard errors. We therefore take empirical results where the robust standard errors fall below the conventional standard errors as a red flag. This is very likely due to bias or a chance occurrence that is better discounted. In this spirit, we like the idea of taking the maximum of the conventional standard error and a robust standard error as your best measure of precision. This rule of thumb helps on two counts: it truncates low values of the robust estimators, reducing bias, and it reduces variability. Table 8.1.1 shows the empirical rejection rates obtained using $Max(HC_j, Conventional)$. The empirical rejection rates using this rule of thumb look pretty good in the first two panels and greatly improve on the robust estimators alone.[8]

Since there is no gain without pain, there must be some cost to using $Max(HC_j, Conventional)$. The cost is that the best standard error when there is no heteroskedasticity is the conventional OLS estimate. This is documented in the bottom panel of the table. Using the maximum inflates standard errors unnecessarily under homoskedasticity, depressing rejection rates. Nevertheless, the table shows that even in this case rejection rates don't go down all that much. We also view an underestimate of precision as being less costly than an over-estimate. Underestimating precision, we come away thinking the data are not very informative and that we should try to collect more data, while in the latter case, we may mistakenly draw important substantive conclusions.

A final comment on this Monte Carlo investigation concerns the sample size. Labor economists like us are used to working with tens of thousands of observations or more. But sometimes we don't. In a study of the effects of busing on public school students, Angrist and Lang (2004) work with samples of about 3000 students grouped in 56 schools. The regressor of interest in this study varies within grade only at the school level, so some of the analysis in this paper uses 56 school means. Not surprisingly, therefore, Angrist and Lang (2004) obtained HC_1 standard errors below conventional OLS standard errors when working with school-level data. As a rule, even if you start with the micro data on individuals, when the regressor of interest varies at a higher level of aggregation - a school, state, or some other group or cluster - effective sample sizes are much closer to the number of clusters than to the number of individuals. Inference procedures for clustered data are discussed in detail in the next section.

[8] Yang, Hsu, and Zhao (2005) formalize the notion of test procedures based on the maximum of a a set of test statistics with differing efficiency and robustness properties.

8.2 Clustering and Serial Correlation in Panels

8.2.1 Clustering and the Moulton Factor

Bias problems aside, heteroskedasticity rarely leads to dramatic changes in inference. In large samples where bias is not likely to be a problem, we might see standard errors increase by about 25 percent when moving from the conventional to the HC_1 estimator. In contrast, clustering can make all the difference.

The clustering problem can be illustrated using a simple bivariate regression estimated in data with a group structure. Suppose we're interested in the bivariate regression,

$$\mathrm{Y}_{ig} = \beta_0 + \beta_1 x_g + e_{ig}, \tag{8.2.1}$$

where Y_{ig} is the dependent variable for individual i in cluster or group g, with G groups. Importantly, the regressor of interest, x_g, varies only at the group level. For example, data from the STAR experiment analyzed by Krueger (1999) come in the form of Y_{ig}, the test score of student i in class g, and class size, x_g.

Although students were randomly assigned to classes in the STAR experiment, the data are unlikely to be independent across observations. The test scores of students in the same class tend to be correlated because students in the same class share background characteristics and are exposed to the same teacher and classroom environment. It's therefore prudent to assume that, for students i and j in the same class, g,

$$E[e_{ig}e_{jg}] = \rho\sigma_e^2 > 0, \tag{8.2.2}$$

where ρ is the intra-class correlation coefficient and σ_e^2 is the residual variance.[9]

Correlation within groups is often modeled using an additive random effects model. Specifically, we assume that the residual, e_{ig}, has a group structure:

$$e_{ig} = v_g + \eta_{ig}. \tag{8.2.3}$$

where v_g is a random component specific to class g and η_{ig} is a mean-zero student-level component that's left over. We focus here on the correlation problem, so both of these error components are assumed to be homoskedastic.

When the regressor of interest varies only at the group level, an error structure like (8.2.3) can increase standard errors sharply. This unfortunate fact is not news - Kloek (1981) and Moulton (1986) both made the point - but it seems fair to say that clustering didn't really become part of the applied econometrics

[9] This sort of residual correlation structure is also a consequence of stratified sampling (see, e.g., Wooldridge, 2003). Most of the samples that we work with are close enough to random that we typically worry more about the dependence due to a group structure than clustering due to stratification.

zeitgeist until about 15 years ago.

Given the error structure, (8.2.3), the intra-class correlation coefficient becomes

$$\rho = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2}.$$

where σ_v^2 is the variance of v_g and σ_η^2 is the variance of η_{ig}. A word on terminology: ρ is called the *intra-class correlation coefficient* even when the groups of interest are not classrooms.

Let $V_c(\widehat{\beta}_1)$ be the conventional OLS variance formula for the regression slope (generated using Ω_c in the previous section), while $V(\widehat{\beta}_1)$ denotes the correct sampling variance given the error structure, (8.2.3). With regressors fixed at the group level and groups of equal size, n, we have

$$\frac{V(\widehat{\beta}_1)}{V_c(\widehat{\beta}_1)} = 1 + (n-1)\rho, \tag{8.2.4}$$

a formula derived in the appendix to this chapter. We call the square root of this ratio the Moulton factor, after Moulton's (1986) influential study. Equation (8.2.4) tells us how much we over-estimate precision by ignoring intra-class correlation. Conventional standard errors become increasingly misleading as n and ρ increase. Suppose, for example, that $\rho = 1$. In this case, all the errors within a group are the same, so the Y_{ig}'s are the same as well. Making a data set larger by copying a smaller one n times generates no new information. The variance $V_c(\widehat{\beta}_1)$ should therefore be scaled up by a factor of n. The Moulton factor increases with group size because with a fixed overall sample size, larger groups means fewer clusters, in which case there is less independent information in the sample (because the data are independent across clusters but not within).[10]

Even small intra-class correlation coefficients can generate a big Moulton factor. In Angrist and Lavy (2007), for example, 4000 students are grouped in 40 schools, so the average n is 100. The regressor of interest is school-level treatment status - all students in treated schools were eligible to receive cash rewards for passing their matriculation exams. The intra-class correlation in this study fluctuates around .1. Applying formula (8.2.4), the Moulton factor is over 3: the standard errors reported by default are only one-third of what they should be.

Equation (8.2.4) covers an important special case where the regressors are fixed within groups and group size is constant. The general formula allows the regressor, x_{ig}, to vary at the individual level and for different group sizes, n_g. In this case, the Moulton factor is the square root of

$$\frac{V(\widehat{\beta}_1)}{V_c(\widehat{\beta}_1)} = 1 + \left[\frac{V(n_g)}{\overline{n}} + \overline{n} - 1\right]\rho_x\rho, \tag{8.2.5}$$

[10] With non-stochastic regressors and homoscedastic residuals, the Moulton factor is a finite-sample result. Survey statisticians call the Moulton factor the *design effect* because it tells us how much to adjust standard errors in stratified samples for deviations from simple random sampling (Kish, 1965).

where $\overline{n}$ is the average group size, and ρ_x is the intra-class correlation of x_{ig}:

$$\rho_x = \frac{\sum_g \sum_{i \neq k} (x_{ig} - \overline{x})(x_{kg} - \overline{x})}{V(x_{ig}) \sum_g n_g(n_g - 1)}.$$

Note that ρ_x does not impose a variance-components structure like (8.2.3) - here, ρ_x is a generic measure of the correlation of regressors within groups. The general Moulton formula tells us that clustering has a bigger impact on standard errors with variable group sizes and when ρ_x is large. The impact vanishes when $\rho_x = 0$. In other words, if the x_{ig}'s are uncorrelated within groups, the grouped error structure does not matter for the estimation of standard errors. That's why we worry most about clustering when the regressor of interest is fixed within groups.

We illustrate formula (8.2.1) using the Tennessee STAR example. A regression of Kindergartners' percentile score on class size yields an estimate of -0.62 with a robust (HC_1) standard error of 0.09. In this case, $\rho_x = 1$ because class size is fixed within classes while $V(n_g)$ is positive because classes vary in size (in this case, $V(n_g) = 17.1$). The intra-class correlation coefficient for residuals is .31 and the average class size is 19.4. Plugging these numbers into (8.2.1) gives a value of about 7 for $\frac{V(\widehat{\beta}_1)}{V_c(\widehat{\beta}_1)}$, so that conventional standard errors should be multiplied by a factor of $2.65 = \sqrt{7}$. The corrected standard error is therefore about 0.24.

The Moulton factor works similarly with 2SLS except that ρ_x should be computed for the instrumental variable and not the regressor. In particular, use (8.2.5) replacing ρ_x with ρ_z, where ρ_z is the intra-class correlation coefficient of the instrumental variable (Shore-Sheppard, 1996) and ρ is the intra-class correlation of the second-stage residuals. To understand why this works, recall that conventional standard errors for 2SLS are derived from the residual variance of the second-stage equation divided by the variance of the first-stage fitted values. This is the same asymptotic variance formula as for OLS, with first-stage fitted values playing the role of regressor.[11]

Here are some solutions to the Moulton problem:

1. Parametric: Fix conventional standard errors using (8.2.5). The intra-class correlations ρ and ρ_x are easy to compute and supplied as descriptive statistics in some software packages.[12]

2. Cluster standard errors: Liang and Zeger (1986) generalize the White (1980a) robust covariance matrix

[11] Clustering can also be a problem in regression-discontinuity designs if the variable that determines treatment assignment varies only at a group level (see Card and Lee, 2008, for details).

[12] Use Stata's `loneway` command, for example.

to allow for clustering as well as heteroskedasticity:

$$\hat{V}_c(\widehat{\beta}) = (X'X)^{-1}\left(\sum_g X_g\widehat{\Psi}_g X_g\right)(X'X)^{-1}, \text{ where} \tag{8.2.6}$$

$$\widehat{\Psi}_g = a\widehat{e}_g\widehat{e}_g' = a\begin{bmatrix} \widehat{e}_{1g}^2 & \widehat{e}_{1g}\widehat{e}_{2g} & \cdots & \widehat{e}_{1g}\widehat{e}_{n_g g} \\ \widehat{e}_{1g}\widehat{e}_{2g} & \widehat{e}_{2g}^2 & & \vdots \\ \vdots & & \ddots & \widehat{e}_{(n_g-1)g}\widehat{e}_{n_g g} \\ \widehat{e}_{1g}\widehat{e}_{n_g g} & \cdots & \widehat{e}_{(n_g-1)g}\widehat{e}_{n_g g} & \widehat{e}_{n_g g}^2 \end{bmatrix}.$$

Here, X_g is the matrix of regressors for group g and a is a degrees of freedom adjustment factor similar to that which appears in HC_1. The clustered variance estimator $\hat{V}_c(\widehat{\beta})$ is consistent as the number of groups gets large under any within-group correlation structure and not just the parametric model in (8.2.3). $\hat{V}_c(\widehat{\beta})$ is not consistent with a fixed number of groups, however, even when the group size tends to infinity. To see why, note that the sums in $\hat{V}_c(\widehat{\beta})$ are over g and not i. Consistency is determined by the law of large numbers, which says that we can rely on sample moments to converge to population moments (Section 3.1.3). But here the sums are at the group level and not over individuals. Clustered standard errors are therefore unlikely to be reliable with few clusters, a point we return to below.

3. Use group averages instead of micro data: let $\overline{\mathrm{Y}}_g$ be the mean of Y_{ig} in group g. Estimate

$$\overline{\mathrm{Y}}_g = \beta_0 + \beta_1 x_g + \overline{e}_g$$

by weighted least squares using the group size as weights. This is equivalent to OLS using micro data but the standard errors are asymptotically correct given the group structure, (8.2.3). Again, the asymptotics here are based on the number of groups and not the group size. Importantly, however, because the group means are close to Normally distributed with modest group sizes, we can expect the good finite-sample properties of regression with Normal errors to kick in. The standard errors that come out of grouped estimation are therefore likely to be more reliable than clustered standard errors in samples with few clusters.

Grouped-data estimation can be generalized to models with micro covariates using a two-step procedure. Suppose the equation of interest is

$$\mathrm{Y}_{ig} = \beta_0 + \beta_1 x_g + \mathrm{W}_{ig}'\delta + e_{ig}, \tag{8.2.7}$$

where W_{ig}' is a vector of covariates that varies within groups. In step 1, construct the covariate-adjusted group effects, μ_g, by estimating

$$\mathrm{Y}_{ig} = \mu_g + \mathrm{W}_{ig}'\delta + \eta_{ig}.$$

The μ_g, called group effects, are coefficients on a full set of group dummies. The estimated $\hat{\mu}_g$ are group means adjusted for the effect of the individual level variables W'_{ig}. Note that by virtue of (8.2.7) and (8.2.3), $\mu_g = \beta_0 + \beta_1 x_g + \nu_g$. In step 2, therefore, we regress the estimated group effects on group-level variables:

$$\widehat{\mu}_g = \beta_0 + \beta_1 x_g + \{\nu_g + (\hat{\mu}_g - \mu_g)\}. \tag{8.2.8}$$

The efficient GLS estimator for (8.2.8) is weighted least squares, using the reciprocal of the estimated variance of the group-level residual, $\{\nu_g + (\hat{\mu}_g - \mu_g)\}$, as weights. This can be a problem since the variance of ν_g is not estimated very well with few groups. We might therefore weight by the reciprocal of the variance of the estimated group effects, the group size, or use no weights at all.[13] In an effort to better approximate the relevant finite-sample distribution, Donald and Lang (2007) suggest that inferences in grouped procedures be based on a t-distribution with $G-\mathrm{K}$ degrees of freedom.

Note that the grouping approach does not work when x_{ig} varies within groups. Averaging x_{ig} to $\bar{x}_g$ is a version of IV, as we saw in Section 4. So with micro-variation in the regressor of interest, grouping estimates parameters that differ from the target parameters in a model like (8.2.7).

4. Block bootstrap: In general, bootstrap inference uses the empirical distribution of the data by re-sampling. But simple random resampling won't do in this case. The trick with clustered data is to preserve the dependence structure in the target population. We do this by block bootstrapping - that is, drawing blocks of data defined by the groups g. In the Tennessee STAR data, for example, we'd block bootstrap by re-sampling entire classes instead of individual students.

5. Estimate a parametric GLS or maximum likelihood model based on a version of (8.2.1). This fixes the clustering problem but also changes the estimand unless the CEF is linear, as detailed in section 3.4.1. We therefore prefer other approaches.

Table 8.2.1 compare standard-error fix-ups in the STAR example. The table reports six estimates of the standard errors: conventional robust standard errors (using HC_1); two versions of parametrically corrected standard errors using the Moulton formula (8.2.5), the first using the formula for the intra-class correlation given by Moulton and the second using Stata's estimator from the *loneway* command; clustered standard errors; block-bootstrapped standard errors; and standard errors from weighted estimation at the group level. The coefficient estimate is -0.62. In this case, all adjustments deliver similar results, a standard error of about .23. This happy outcome is due in large part to the fact that with 318 classrooms, we have enough clusters for group-level asymptotics to work well. With few clusters, however, things are much dicier, a point we return to at the end of the chapter.

[13] See, e.g., Angrist and Lavy (2007) for an example of the latter two weighting schemes.

8.2.2 Serial Correlation in Panels and Difference-in-Difference Models

Serial correlation - the tendency for one observation to be correlated with those that have gone before - used to be Somebody Else's Problem, specifically, the unfortunate souls who make their living out of time series data (macroeconomists, for example). Applied microeconometricians have therefore long ignored it.[14] But our data often have a time dimension too, especially in differences-in-differences models. This fact combined with clustering can have a major impact on statistical inference.

Suppose, as in Section 5.2, that we are interested in the effects of a state minimum wage. In this context, the regression version of differences-in-differences includes additive state and time effects. We therefore we get an equation like (5.2.3), repeated below:

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \beta \mathrm{D}_{st} + \varepsilon_{ist}, \tag{8.2.9}$$

As before, Y_{ist} is the outcome for individual i in state s in year t and D_{st} is a dummy variable that indicates treatment states in post-treatment periods.

The error term in (8.2.9) reflects the idiosyncratic variation in potential outcomes that varies across people, states, and time. Some of this variation is likely to be common to individuals in the same state and year, for example, a regional business cycle. We can model this common component by thinking of ε_{ist} as the sum of a state-year shock, v_{st}, and an idiosyncratic individual component, η_{ist}. So we have:

$$\mathrm{Y}_{ist} = \gamma_s + \lambda_t + \beta \mathrm{D}_{st} + v_{st} + \eta_{ist}. \tag{8.2.10}$$

We assume that in repeated draws across states and over time, $E[v_{st}] = 0$, while $E[\eta_{ist}] = 0$ by definition.

State-year shocks are bad news for differences-in-differences models. As with the Moulton problem, state- and time-specific random effects generate a clustering problem that affects statistical inference. But that might be the least of our problems in this case. To see why, suppose we have only two periods and two states, as in the Card and Krueger (1994) New Jersey/Pennsylvania study. The empirical difference-in-differences is

$$\hat{\beta}_{CK} = (\overline{\mathrm{Y}}_{s=NJ,t=Nov} - \overline{\mathrm{Y}}_{s=NJ,t=Feb}) - (\overline{\mathrm{Y}}_{s=PA,t=Nov} - \overline{\mathrm{Y}}_{s=PA,t=Feb}).$$

This estimator is unbiased since $E[v_{st}] = E[\eta_{ist}] = 0$. On the other hand, assuming we think of probability limits as increasing group size while keeping the choice of states and periods fixed, state-year shocks render

[14] The *Somebody Else's Problem (SEP) Field*, first identified as a natural phenomenon in Adams' *Life, the Universe, and Everything, i*s, according to Wikipedia, "a generated energy field that affects perception . . . Entities within the field will be perceived by an outside observer as 'Somebody Else's Problem', and will therefore be effectively invisible unless the observer is specifically looking for the entity."

$\hat{\beta}_{CK}$ inconsistent:

$$plim\hat{\beta}_{CK} = \beta + \{(v_{s=NJ,t=Nov} - v_{s=NJ,t=Feb}) - (v_{s=PA,t=Nov} - v_{s=PA,t=Feb})\}.$$

Averaging larger and larger samples within New Jersey or Pennsylvania in a given period does nothing to eliminate the regional shocks specific to a given location and period. With only two states and years, we have no way to distinguish the differences-in-differences generated by a policy change from the difference-in-dfferences due to the fact that, say, the New Jersey economy was holding steady in 1992 while Pennsylvania was experiencing a mild cyclical downturn. We can think of the presence of v_{st} as a failure of the common trends assumption discussed in Section 5.2.

The solution to the inconsistency induced by random shocks in differences in differences models is to have either multiple time periods or many states (or both). For example, Card (1992) uses 51 states to study minimum wage changes while Card and Krueger (2000) take another look at the New Jersey-Pennsylvania experiment with a longer monthly time series of payroll data. With multiple states and/or periods, we can hope that the v_{st} average out to zero. As in the first part of this chapter on the Moulton problem, the inference framework in this context relies on asymptotic distribution theory with many groups and not on group size (or, at least, not on group size alone). The most important inference issue then becomes the behavior of v_{st}. In particular, if we are prepared to assume that shocks are independent across states and over time - i.e., they are serially uncorrelated - we are back to the plain-vanilla Moulton problem in Section 8.2.1, in which case we would cluster by state $\times$ year. But in most cases, the assumption that v_{st} is serially uncorrelated is hard to defend. Almost certainly, for example, regional shocks are highly serially correlated: if things are bad in Pennsylvania in one month, they are likely to be just about as bad in the next.

The consequences of serial correlation for clustered panels are highlighted by Bertrand, Duflo, and Mullainathan (2004) and Kézdi (2004). Any research design with a group structure where the group means are correlated can be said to have the serial correlation problem. The upshot of recent work on serial correlation in data with a group structure is that, just as we must adjust our standard errors for the correlation *within* groups induced by the presence of v_{st}, we must further adjust for serial correlation in the v_{st} themselves. There are a number of ways to do this, not all equally effective in all situations. It seems fair to say that the question of how best to approach the serial correlation problem is currently under study and a consensus has not yet emerged. We try here to give a flavor of the approaches and summarize the emerging findings.

The simplest and most widely applied approach is simply to pass the clustering buck one level higher. So in the state-year example, we can report Liang and Zeger (1986) standard errors clustered by state instead of by state and year (e.g., using Stata cluster). This might seem odd at first blush, since the model controls for state effects. The state effect, γ_s, in (8.2.10) removes the time mean of v_{st}, which we denote by $\overline{v}_s$. Nevertheless, $v_{st} - \overline{v}_s$ is probably still serially correlated. Clustering at the state level takes account

of this since the one-level-up clustered covariance estimator allows for completely non-parametric residual correlation within clusters - including the time series correlation in $v_{st} - \overline{v}_s$. This is a quick and easy fix. The problem here, as you might have guessed, is that passing the buck up one level reduces the number of clusters. And asymptotic inference supposes we have a large number of clusters because we need a lot of states or periods to estimate the correlation between $v_{st} - \overline{v}_s$ and $v_{st-1} - \overline{v}_s$ reasonably well. Few clusters means biased standard errors and misleading inferences.

8.2.3 Fewer than 42 clusters

Bias from few clusters is a risk in both the Moulton and the serial correlation contexts because in both cases inference is cluster-based. With few clusters, we tend to underestimate either the serial correlation in a random shock like v_{st} or the intra-class correlation, ρ, in the Moulton problem. The relevant dimension for counting clusters in the Moulton problem is the number of groups, G. In a differences-in-differences scenario where you'd like to cluster on state (or some other cross-sectional dimension), the relevant dimension for counting clusters is the number of states or cross-sectional groups. Therefore, following Douglas Adam's dictum that the ultimate answer to life, the universe, and everything is 42, we believe the question is: How many clusters are enough for reliable inference using a standard cluster adjustment derived from (8.2.6)?

If 42 is enough for the standard cluster adjustment to be reliable - and less is too few - then what should you do when the cluster count is low? First-best is to get more clusters by collecting more data. But sometimes we're too lazy for that, so other ideas are detailed below. It's worth noting at the outset that not all of these ideas are equally well-suited for the Moulton and serial correlation problems.

1. Bias correction of clustered standard errors. Clustered standard errors are biased in small samples because $E\left(\widehat{e}_g\widehat{e}_g'\right) \neq E\left(e_g e_g'\right) = \Psi_g$ just as in Section 8.1. Usually, $E\left(\widehat{e}_g\widehat{e}_g'\right)$ is too small. One solution is to inflate residuals in the hopes of reducing bias. Bell and McCaffrey (2002) suggest a procedure (called bias-reduced linearization or BRL) that adjusts residuals by

$$\begin{aligned} \hat{\Psi}_g &= a\widetilde{e}_g\widetilde{e}_g' \\ \widetilde{e}_g &= A\widehat{e}_g \end{aligned}$$

where A solves

$$A_g'A_g = (I - H_g)^{-1}$$

and

$$H_g = X_g(X'X)^{-1}X_g'.$$

This is a version of HC_2 for the clustered case. BRL works for the straight-up Moulton problem with few clusters but for technical reasons cannot be used for the typical differences-in-differences serial

correlation problem.[15]

2. Recognizing that the fundamental unit of observation is a cluster and not an individual unit within clusters, Bell and McCaffrey (2002) and Donald and Lang (2007) suggest that inference be based on a t-distribution with $G-\text{K}$ degrees of freedom rather than on the standard Normal distribution. For small G, this makes a big difference - confidence intervals will be much wider, thereby avoiding some mistakes. Cameron, Gelbach, and Miller (2008) report Monte Carlo examples where the combination of a BRL adjustment and use of t-tables works well.

3. Donald and Lang (2007) argue that estimation using group means works well with small G in the Moulton problem, and even better when inference is based on a t-distribution with $G-\text{K}$ degrees of freedom. But, as we discussed in the previous section, the regressor must be fixed within groups. The level of aggregation is the level at which you'd like to cluster, e.g., schools in Angrist and Lavy (2007). For serial correlation, this is the state, but state averages cannot be used to estimate a model with a full set of state effects. Also, since treatment status varies within states, averaging up to the state level averages the regressor of interest as well, changing the rules of the game in a way we may not like (the estimator becomes instrumental variables using group dummies as instruments). The group means approach is therefore out of bounds for the serial correlation problem.[16] Note also that if the grouped residuals are heteroskedastic, and you therefore use robust standard errors, you must worry about bias of the form discussed in Section 8.1. If both the random effect and the underlying micro residual are homoskedastic, you can fix heteroskedasticity in the group means by weighting by the group size. But weighting changes the estimand when the CEF is nonlinear - so this is not open-and-shut (Angrist and Lavy, 1999 chose not to weight school-level averages because the variation in their study comes mostly from small schools). Weighted or not, the safest course when working with group-level averages is to use of our rule of thumb from Section 8.1: take the maximum of robust and conventional standard errors as your best measure of precision.

[15] The matrix A_g is not unique; there are many such decompositions. Bell and McCaffrey (2002) use the symemtric square root of $(I - H_g)^{-1}$ or

$$A_g = P\Lambda^{1/2}$$

where P is the matrix of eigenvectors of $(I - H_g)^{-1}$, Λ is the diagonal matrix of the correponding eigenvalues, and $\Lambda^{1/2}$ is the diagonal matrix of the square roots of the eigenvalues. One problem with the Bell and McCaffrey adjustment is that $(I - H_g)$ may not be of full rank, and hence the inverse may not exist for all designs. This happens, for example, when one of the regressors is a dummy variable which is one for exactly one of the clusters, and zero otherwise. This includes the panel DD model discussed by Bertrand et al. (2004), where you include a full set of state dummies and cluster by state. Moreover, the eigenvalue decomposition is implemented for matrices which are the size of the groups. In many applications, group sizes are large enough that this becomes computationally intractible.

[16] Donald and Lang (2007) discuss serial correlation examples where the regressor is fixed within the clustering dimension, but this is not the typical differences-in-differences setup.

4. Cameron, Gelbach, and Miller (2008) report that some forms of a block bootstrap work well with small numbers of groups, and that the block bootstrap typically outperforms Stata-clustered standard errors without the bias correction. This appears to be true both for the Moulton and serial correlation problems. But Cameron, Gelbach, and Miller (2008) focus on rejection rates using (pivotal) test statistics, while we like to see standard errors.

5. Parametric corrections: For the Moulton problem, this amounts to use of the Moulton factor. With serial correlation, this means correcting your standard errors for first-order serial correlation at the group level. Based on our sampling experiments with the Moulton problem and a reading of the literature, parametric approaches may work well, and better than the nonparametric estimator (8.2.6), especially if the parametric model is not too far off (see, e.g., Hansen, 2007a, which also proposes a bias correction for estimates of serial correlation parameters). Unfortunately, however, beyond the greenhouse world of controlled Monte Carlo studies, we're unlikely to know whether parametric assumptions are a good fit.

Alas, the bottom line here is not entirely clear, as is the more basic question of when few clusters are fatal for inference. The severity of the resulting bias seems to depend on the nature of your problem, in particular whether you confront straight-up Moulton or serial correlation issues. Aggregation to the group level as in Donald and Lang (2007) seems to work well in the Moulton case as long as the regressor of interest is fixed within groups and there is not too much underlying heteroskedasticity. At a minimum, you'd like to show that your conclusions are consistent with the inferences that arise from an analysis of group averages since this is a conservative and transparent approach. Angrist and Lavy (2007) go with BRL standard errors to adjust for clustering at the school level but validate these by showing that key results come out the same using covariate-adjusted group averages.

As far as serial correlation goes, most of the evidence suggests that when you are lucky enough to do research on US states, giving 51 clusters, you are on reasonably safe ground with a naive application of Stata's cluster command at the state level. But you might have to study Canada, which offers only 10 clusters in the form of provinces, well below 42. Hansen (2007b) finds that Liang and Zeger (1986) [Stata-clustered] standard errors are reasonably good at correcting for serial correlation in panels, even in the Canadian scenario. Hansen also recommends use of a t-distribution with $G - k$ degrees of freedom for critical values.

Clustering problems have forced applied microeconometricians to eat a little humble pie. Proud of working with large micro data sets, we like to sneer at macroeconomists toying with small time series samples. But he who laughs last laughs best: if the regressor of interest varies only at a coarse group level - such as over time or across states or countries - then it's the macroeconomists who have had the most realistic mode of inference all along.

8.3 Appendix: Derivation of the simple Moulton factor

Write

$$y_g = \begin{bmatrix} y_{1g} \\ y_{2g} \\ \vdots \\ y_{n_g g} \end{bmatrix} \quad e_g = \begin{bmatrix} e_{1g} \\ e_{2g} \\ \vdots \\ e_{n_g g} \end{bmatrix}$$

and

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix} \quad x = \begin{bmatrix} \iota_1 x_1 \\ \iota_2 x_2 \\ \vdots \\ \iota_G x_G \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_G \end{bmatrix}$$

where ι_g is a column vector of n_g ones and G is the number of groups. Note that

$$\begin{aligned} E(ee') &= \Psi = \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \Psi_G \end{bmatrix} \\ \Psi_g &= \sigma_e^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix} = \sigma_e^2 \left[(1-\rho) I + \rho \iota_g \iota_g' \right], \\ \text{where } \rho &= \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2}. \end{aligned}$$

Now

$$\begin{aligned} X'X &= \sum_g n_g x_g x_g' \\ X'\Psi X &= \sum_g x_g \iota_g' \Psi_g \iota_g x_g'. \end{aligned}$$

But

$$\begin{aligned} x_g \iota_g' \Psi_g \iota_g x_g' &= \sigma_e^2 x_g \iota_g' \begin{bmatrix} 1+(n_g-1)\rho \\ 1+(n_g-1)\rho \\ \cdots \\ 1+(n_g-1)\rho \end{bmatrix} x_g' \\ &= \sigma_e^2 n_g \left[1+(n_g-1)\rho \right] x_g x_g'. \end{aligned}$$

Let $\tau_g = 1 + (n_g - 1)\rho$, so we get

$$\begin{aligned} x_g \iota_g' \Psi_g \iota_g x_g' &= \sigma_e^2 n_g \tau_g x_g x_g' \\ X'\Psi X &= \sigma_e^2 \sum_g n_g \tau_g x_g x_g'. \end{aligned}$$

With this in hand, we can write

$$\begin{aligned} V(\widehat{\beta}) &= (X'X)^{-1} X'\Psi X (X'X)^{-1} \\ &= \sigma_e^2 \left(\sum_g n_g x_g x_g' \right)^{-1} \sum_g n_g \tau_g x_g x_g' \left(\sum_g n_g x_g x_g' \right)^{-1}. \end{aligned}$$

We want to compare this with the standard OLS covariance estimator

$$V_c(\widehat{\beta}) = \sigma_e^2 \left(\sum_g n_g x_g x_g' \right)^{-1}.$$

If the group sizes are equal, $n_g = n$ and $\tau_g = \tau = 1 + (n-1)\rho$, so that

$$\begin{aligned} V(\widehat{\beta}) &= \sigma_e^2 \tau \left(\sum_g n x_g x_g' \right)^{-1} \sum_g n x_g x_g' \left(\sum_g n x_g x_g' \right)^{-1} \\ &= \sigma_e^2 \tau \left(\sum_g n x_g x_g' \right)^{-1} \\ &= \tau V_c(\widehat{\beta}), \end{aligned}$$

which implies (8.2.4).

Table 8.1.1: Monte Carlo results for robust standard errors

	Mean	Standard Deviation	Empirical 5% Rejection Rates: Normal	Empirical 5% Rejection Rates: t
	(1)	(2)	(3)	(4)
A. Lots of Heteroskedasticity				
$\hat{\beta}_1$	-0.001	0.586		
Standard Errors:				
Conventional	0.331	0.052	0.278	0.257
HC0	0.417	0.203	0.247	0.231
HC1	0.447	0.218	0.223	0.208
HC2	0.523	0.26	0.177	0.164
HC3	0.636	0.321	0.13	0.12
max(Conventional, HC0)	0.448	0.172	0.188	0.171
max(Conventional, HC1)	0.473	0.19	0.173	0.157
max(Conventional, HC2)	0.542	0.238	0.141	0.128
max(Conventional, HC3)	0.649	0.305	0.107	0.097
B. Little Heteroskedasticity				
$\hat{\beta}_1$	0.004	0.6		
Standard Errors:				
Conventional	0.52	0.07	0.098	0.084
HC0	0.441	0.193	0.217	0.202
HC1	0.473	0.207	0.194	0.179
HC2	0.546	0.25	0.156	0.143
HC3	0.657	0.312	0.114	0.104
max(Conventional, HC0)	0.562	0.121	0.083	0.07
max(Conventional, HC1)	0.578	0.138	0.078	0.067
max(Conventional, HC2)	0.627	0.186	0.067	0.057
max(Conventional, HC3)	0.713	0.259	0.053	0.045
C. No Heteroskedasticity				
$\hat{\beta}_1$	-0.003	0.611		
Standard Errors:				
Conventional	0.604	0.081	0.061	0.05
HC0	0.453	0.19	0.209	0.193
HC1	0.486	0.203	0.185	0.171
HC2	0.557	0.247	0.15	0.136
HC3	0.667	0.309	0.11	0.1
max(Conventional, HC0)	0.629	0.109	0.055	0.045
max(Conventional, HC1)	0.64	0.122	0.053	0.044
max(Conventional, HC2)	0.679	0.166	0.047	0.039
max(Conventional, HC3)	0.754	0.237	0.039	0.031

Note: The table reports results from a sampling experiment with 25,000 replications.

Table 8.2.1: Standard errors for class size effects in the STAR data

	Standard Error
Robust (HC1)	0.09
Parametric Moulton Correction (using Moulton intraclass coefficient)	0.222
Parametric Moulton Correction (using ANOVA intraclass coefficient)	0.23
Clustered	0.232
Block Bootstrap	0.231
Estimation using group means (weighted by class size)	0.226

Note: The table reports estimates from a regression of average percentile scores on class size for kindergartners using the public use data set from Project STAR. The coefficient on class size is -.62. The group level for clustering is the classroom. The number of observations is 5,743. The bootstrap estimate uses 1,000 replications.

Last words

If applied econometrics was easy, theorists would do it. But it's not as hard as the dense pages of *Econometrica* might lead you to believe. Carefully applied to coherent causal questions, regression and 2SLS almost always make sense. Your standard errors probably won't be quite right, but they rarely are. Avoid embarrassment by being your own best skeptic - and, especially, Don't Panic!

Acronyms

Technical terms

2SLS Two Stage Least Squares, an Instrumental Variables(IV) estimator (89)

ACR Average Causal Response, the weighted average causal response to an ordered treatment (136)

ANOVA Analysis of Variance, a decomposition of total variance into the variance of the Conditional Expectation Function (CEF) and the average conditional variance (26)

BRL Biased Reduced Linearization estimator, a bias-corrected covariance matrix estimator for clustered data (238)

CDF Cumulative Distribution Function, the probability that a random variable takes on a value less than or equal to a given number (72)

CEF Conditional Expectation Function, the population average of Y_i with X_i held fixed (23)

CIA Conditional Independence Assumption, a core assumption that justifies a causal interpretation of regression and matching estimators (39)

COP Conditional on Positive effect, the treatment-control difference in means for a non-negative random variable looking at positive values only (73)

CQF Conditional Quantile Function, defined for each quantile τ, the $\tau-$quantile of Y_i holding X_i fixed (204)

DD Differences in Differences estimator, in it's simplest form, a comparison of changes over time in treatment and control groups (169)

GLS Generalized Least Squares estimator, a regression estimator for models with heteroskedasticity and/or serial correlation; GLS provides efficiency gains when the Conditional Expectation Function (CEF) is linear (69)

GMM Generalized Method of Moments, an econometric estimation framework in which estimates are chosen to minimize a matrix-weighted average of the squared difference between sample and population moments (105)

HC_0 **-** HC_3 Heteroskedasticity Consistent variance estimators proposed by MacKinnon and White (1985) (227)

ILS Indirect Least Squares estimator, the ratio of reduced form to first-stage coefficients in an Instrumental Variables (IV) set-up (89)

ITT Intention to Treat effect, the effect of being offered treatment (122)

IV Instrumental Variables estimator (83)

LATE Local Average Treatment Effect, the causal effect of treatment on compliers (114)

LDVs Limited Dependent Variables, e.g., dummies, counts, and non-negative random variables on the left-hand side of regression and related statistical models (70)

LIML Limited Information Maximum Likelihood estimator, an alternative to Two-Stage Least Squares (2SLS) with less bias (109)

LM Lagrange Multiplier test, a statistical test of the restrictions imposed by an estimator (108)

LPM Linear Probability Model (36)

MFX Marginal Effects, in nonlinear models, the derivative of the Conditional Expectation Function (CEF) implied by the model with respect to the regressors (78)

MMSE Minimum Mean Squared Error, minimum expected squared prediction error, or the minimum of the expected square of the difference between an estimator and a target (25)

OLS Ordinary Least Squares estimator, the sample analog of the population regression vector (78)

OVB Omitted Variables Bias formula, the relationship between regression estimates in models with different sets of control variables (44)

QTE Quantile Treatment Effect, the causal effect of treatment on conditional quantiles of the outcome variable for compliers (215)

RD Regression Discontinuity design, an identification strategy in which treatment, the probability of treatment, or the average treatment intensity is a known, discontinuous function of a covariate (189)

SEM Simultaneous Equations Models, an econometric framework in which causal relationships between variables are described by several equations (84)

SSIV Split-Sample Instrumental Variables estimator, a version of the Two-Sample Instrumental Variables (TSIV) estimator (111)

TSIV Two-Sample Instrumental Variables estimator, an Instrumental Variables (IV) estimator that can sometimes be constructed from two data sets, when either data set alone would be inadequate (109)

VIV Visual Instrumental Variables, a plot of reduced-form against first-stage fitted values in instrumental variables models with dummy instruments (103)

Data sets and variable names

AFDC Aid to Families with Dependent Children, an American welfare program no longer in effect (121)

AFQT Armed Forces Qualification Test, used by the US armed forces to gauge recruits' academic and cognitive ability (46)

CPS Current Population Survey, a large monthly survey of US households, source of the US unemployment rate (45)

GED General Educational Development certificate, a substitute for traditional high school credentials, obtained by passing a test (121)

IPUMS Integrated Public Use Microdata Series, consistently coded samples of census records from the US and other countries (24)

NHIS National Health Interview Survey, a large American survey with many questions related to health (10)

NLSY National Longitudinal Survey of Youth, a long-running panel survey that started with a high-school-aged cohort (46)

PSAT Preliminary SAT, qualifies American high school sophomores for a National Merit Scholarship (189)

PSID Panel Study of Income Dynamics, a panel survey of American households begun in 1968 (64)

QOB Quarter of Birth (92)

RSN Random Sequence Numbers, draft lottery numbers randomly assigned to dates of birth in the Vietnam-era draft lotteries held from 1970-73 (95)

SDA Service Delivery Area, one of the 649 sites where Job Training Partnership Act(JTPA) services were delivered (217)

SSA Social Security Administration (110)

Study Names

HIE Health Insurance Experiment conducted by the RAND Corporation, a randomized trial in which participants were exposed to insurance programs with different features (70)

HRT Hormone Replacement Therapy, an intervention designed to reduce the symptoms of menopause (12)

JSA Job Search Assistance, part of the Job Training Partnership Act (JTPA) (218)

JTPA Job Training Partnership Act, a large federal training program which included a randomized evaluation (119)

MDVE Minneapolis Domestic Violence Experiment, a randomized trial in which police response to a domestic disturbance was determined in part by random assignment (123)

NSW National Supported Work demonstration, an experimental mid-1970s training program that provided work experience to a sample with weak labor-force attachment (64)

OJT On the Job Training, part of the Job Training Partnership Act (JTPA) (218)

STAR the Tennessee Student/Teacher Achievement Ratio experiment, a randomized study of elementary school class size (13)

WHI Women's Health Initiative, a series of randomized trials that included an evaluation of Hormone Replacement Therapy (HRT) (12)

Empirical Studies Index

Note: Page numbers below refer to text locations where key elements of the study are described.

Angrist, Chernozhukov, Fernandez-Val (2006) Shows that quantile regression estimates a MMSE approximation to a nonlinear CQF, and illustrates the quantile regression approximation property by estimating the effects of schooling on the distribution of wages. Discussed in Section 7.1.2. Results appear in Table 7.1.1 and Figure 7.1.1.

Autor (2003) Uses state variation in employment protection laws to construct DD estimates of the effect of labor market regulation on temporary employment. Discussed on page 177. Results appear in Figure 5.2.4.

Besley and Burgess (2004) Use state variation to estimate the effect of labor laws on firm performance in India. Discussed on page 178. Results appear in Table 5.2.3.

Card (1992) Uses state minimum wages regional variation in minimum wage laws to estimate the effect of the minimum wage, Discussed on page 175. Results appear in Table 5.2.2.

Card and Krueger (1994, 2000) Uses a New Jersey minimum wage increase to estimate the employment effects of a minimum wage change. Discussed in Section 5.2. Results appear in Table 5.2.1.

Dehejia and Wahba (1999) Uses the propensity score to estimate the effects of subsidized training on earnings, in a re-analysis of the Lalonde (1986) NSW sample. Discussed on page 64. Results appear in Table 3.3.2.

Freeman (1984) Uses fixed effects models to construct panel-data estimates of the effect of union status on wages. Discussed on page 167. Results appear in Table 5.1.1.

Krueger (1999) Uses the Tennessee randomized trial to construct IV estimates of the effect of class size on test scores. Discussed on page 13. Results appear in Table 2.2.1, 2.2.2, and Table 8.2.1.

Lee (2008) Uses a regression discontinuity design to estimate the effect of party incumbency on re-election. Discussed on page 194. Results appear in Figure 6.1.2.

Manning, et al (1987) Uses randomized assignment to estimate the impact of health insurance plans on health care use, cost, and outcomes. Discussed on page 70. Results appear in Table 3.4.1.

Pischke (2007) Uses a sharp change in the length of the German school year to estimate the effect of school term length on achievement. Discussed on page 172. Results appear in Figure 5.2.3.

Notation

X_i	a $\mathrm{K}\times 1$ vector of covariates, with elements x_{ki}; $k = 1, ...,\mathrm{K}$
x_i	the single regressor in a bivariate regression
Y_i	an outcome or dependent variable
ε_i	$\equiv \mathrm{Y}_i - E[\mathrm{Y}_i \vert \mathrm{X}_i]$, the CEF residual
β	$\equiv \arg\min_b E\left[\left(\mathrm{Y}_i - \mathrm{X}_i'b\right)^2\right]$, the population regression vector; $\beta = E\left[\mathrm{X}_i\mathrm{X}_i'\right]^{-1} E\left[\mathrm{X}_i\mathrm{Y}_i\right]$
e_i	$\equiv \mathrm{Y}_i - \mathrm{X}_i'\beta$, a population regression residual
$\tilde{x}_{ki}$	the residual from a regression of regressor x_{ki} on all other covariates in the model
w_i	the inverse probability of sampling observation i
$\hat{\beta}$	$\equiv \left[\sum_i \mathrm{X}_i\mathrm{X}_i'\right]^{-1} \sum_i \mathrm{X}_i\mathrm{Y}_i$, the OLS estimator
$\widehat{e}_i$	$\equiv \mathrm{Y}_i - \mathrm{X}_i'\hat{\beta}$, the estimated residual
$f_i(s)$	an individual-specific causal relationship between schooling and earnings, e.g., the amount i would earn with s years of schooling
δ_{TOT}	$\equiv E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} \vert \mathrm{D}_i = 1]$, the effect of treatment on the treated
δ_{ATE}	$\equiv E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}]$, the average treatment effect
$h(s)$	$\equiv E[\mathrm{Y}_i \vert \mathrm{S}_i = s]$, the CEF of Y_i given schooling equal to s
μ_t	$\equiv \{E[\mathrm{S}_i \vert \mathrm{S}_i \geq t] - E[\mathrm{S}_i \vert \mathrm{S}_i < t]\}\{P(\mathrm{S}_i \geq t)[1 - P(\mathrm{S}_i \geq t)]\}$, the implicit weight on S_i when the population regression of Y_i on S_i is interpreted as a weighted average of $h'(s)$
δ_R	the population regression of Y_i on D_i, controlling for a saturated model for covariates
Y_i^*	a latent outcome variable, related to the observed outcome variable by $\mathrm{Y}_i = 1[\mathrm{Y}_i^* > 0]$
A_i	a vector of omitted variables in a regression (e.g., "ability" in a regression of wages on schooling)
Z_i	a dummy instrumental variable; if more than one instrument, $\mathrm{Z}_{qi}; q = 1, ...,\mathrm{Q}$. In a vector combining instruments with exogenous covariates, $Z_i \equiv \begin{bmatrix} \mathrm{X}_i' & \mathrm{Z}_{1i} & ... & \mathrm{Z}_{\mathrm{Q}i} \end{bmatrix}'$
$\hat{s}_i$	fitted values in a population regression of S_i on covariates and instruments, $\mathrm{X}_i'\hat{\pi}_{10} + \hat{\pi}_{11}\mathrm{Z}_i$
D_{0i},D_{1i}	a pair of potential treatment assignments indexed against Z_i
D_i	the observed treatment variable, equal to $(1-\mathrm{Z}_i)\mathrm{D}_{0i} + \mathrm{Z}_i\mathrm{D}_{1i}$ in an IV set-up

Γ $\equiv \begin{bmatrix} \alpha' & \rho \end{bmatrix}'$, the vector of coefficients in a 2SLS second stage equation, where the coefficient of interest is ρ

$\hat{\Gamma}_{2SLS}$ $\equiv [\sum_i V_i V_i']^{-1} \sum_i V_i \mathrm{Y}_i$, a 2SLS estimator
$= [W' P_Z W]^{-1} W' P_Z \mathrm{Y}$

$\mathrm{Y}_i(d, z)$ the potential outcome of individual i were this person to have treatment status $\mathrm{D}_i = d$ and instrument value $\mathrm{Z}_i = z$.

ρ_i $\equiv \mathrm{Y}_{1i} - \mathrm{Y}_{0i}$, the individual treatment effect in a random coefficients setup with a binary treatment D_i

π_{1i} heterogeneous causal effect of an instrument on D_i in random coefficients setup: $\mathrm{D}_i = \mathrm{D}_{0i} + (\mathrm{D}_{1i} - \mathrm{D}_{0i})\mathrm{Z}_i = \pi_0 + \pi_{1i}\mathrm{Z}_i + v_i$.

κ_i Abadie kappa, $\kappa_i = 1 - \frac{\mathrm{D}_i(1-\mathrm{Z}_i)}{1-P(\mathrm{Z}_i=1|\mathrm{X}_i)} - \frac{(1-\mathrm{D}_i)\mathrm{Z}_i}{P(\mathrm{Z}_i=1|\mathrm{X}_i)}$, the weight used to find the expectation of any function of the data for compliers

η_i error term in a causal model, e.g., $y_i = \beta x_i + \eta_i$

ξ_i error term in a 1st stage regression, e.g., $x_i = Z_i'\pi + \xi_i$

$\varepsilon_{it}, \varepsilon_{ist}$ population regression errors in panel data in chapter 5

$\Phi[\cdot]$ standard normal cumulative distribution function (CDF)

$\phi[\cdot]$ standard normal density

$\Phi_b(\cdot, \cdot; \rho_{\varepsilon\nu})$ bivariate standard normal CDF with correlation coefficient $\rho_{\varepsilon\nu}$

Y_{it-h} observation on the dependent variable h periods ago

Δ difference operator, e.g. $\Delta \mathrm{Y}_{it} = \mathrm{Y}_{it} - \mathrm{Y}_{it-1}$

$F_Y(y|\mathrm{X}_i)$ the distribution function for Y_i conditional on X_i.

$Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$ $\equiv F_Y^{-1}(\tau|\mathrm{X}_i)$, conditional quantile function (CQF)

$\rho_\tau(u)$ $= (\tau - 1(u \leq 0))u$, check function, the expectation of which is minimized by the CQF

β_τ $\equiv \arg\min_b E\left[\rho_\tau(\mathrm{Y}_i - \mathrm{X}_i'b)\right]$, population quantile regression vector

$\Delta_\tau(\mathrm{X}_i, \beta_\tau)$ $\equiv \mathrm{X}_i'\beta_\tau - Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$, quantile regression specification error

Ω asymptotic covariance matrix of the OLS estimator

Ψ $\equiv E[ee']$, variance matrix of residuals, with diagonal elements ψ_i

e_{ig} $\equiv \nu_g + \eta_{ig}$, error with a group structure in chapter 8

$\hat{\Omega}_c$ $= (X'X)^{-1}\left(\sum \frac{\widehat{e}_i^2}{N}\right)$, conventional variance estimator

$\hat{\Omega}_r$ $= (X'X)^{-1}\left(\sum \frac{\mathrm{X}_i\mathrm{X}_i\widehat{e}_i^2}{N}\right)(X'X)^{-1}$, robust variance estimator

H $\equiv X(X'X)^{-1}X'$, covariate projection matrix

h_{ii} $\equiv \mathrm{X}_i'(X'X)^{-1}\mathrm{X}_i$, the *leverage* of the ith observation, the ith diagonal element of H

M $\equiv I_N - H$, the residual maker matrix

References

Abadie, Alberto (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Abadie, Alberto, Joshua D. Angrist, and Guido Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2007): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," National Bureau of Economic Research, Working Paper No. 12831.

Abadie, Alberto, and Guido Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–67.

——— (2008): "Bias-Corrected Matching Estimators for Average Treatment Effects," Harvard University, Department of Economics, mimeo.

Acemoglu, Daron, and Joshua Angrist (2000): "How Large are the Social Returns to Education? Evidence from Compulsory Schooling Laws," in *National Bureau of Economics Macroeconomics Annual 2000*, ed. by Ben S. Bernanke, and Kenneth S. Rogoff, pp. 9–58. The MIT Press, Cambridge.

Acemoglu, Daron, Simon Johnson, and James A. Robinson (2001): "The Colonial Origins of Comparative Development: An Empirical Investigation," *The American Economic Review*, 91, 1369–1401.

Adams, Douglas (1979): *The Hitchhiker's Guide to the Galaxy*. Pocket Books, New York.

——— (1990): *Dirk Gently's Holistic Detective Agency*. Simon & Schuster, New York.

——— (1995): *Mostly Harmless*. Harmony Books, New York.

Altonji, Joseph G., and Lewis M. Segal (1996): "Small-Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics*, 14, 353–366.

Ammermueller, Andreas, and Jorn-Steffan Pischke (2006): "Peer Effects in European Primary Schools: Evidence from PIRLS," Institute for the Study of Labor (IZA), Discussion Paper No. 2077.

Ananat, Elizabeth, and Guy Michaels (2008): "The Effect of Marital Breakup on the Income Distribution of Women with Children," *Journal of Human Resources*, forthcoming.

Anderson, Michael (2008): "Multiple Inference and Gender Differences in the Effect of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, forthcoming.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer (2002): "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment," *The American Economic Review*, 92, 1535–1558.

Angrist, Joshua D. (1988): "Grouped Data Estimation and Testing in Simple Labor Supply Models," Princeton University, Industrial Section, Working Paper No. 234.

——— (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313–335.

——— (1991): "Grouped Data Estimation and Testing in Simple Labor Supply Models," *Journal of Econometrics*, 47, 243–266.

——— (1998): "Estimating the Labor Market Impact on Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66, 249–288.

——— (2001): "Estimations of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics*, 19, 2–16.

——— (2004): "American Education Research Changes Track," *Oxford Review of Economic Policy*, 20, 198–212.

——— (2006): "Instrumental Variables Methods in Experimental Criminological Research: What, Why and How," *Journal of Experimental Criminology*, 2, 22–44.

Angrist, Joshua D., Victor Chernozhukov, and Ivan Fernandez-Val (2006): "Quantile Regression Under Misspecification, with an Application to the U.S. Wage Structure," *Econometrica*, 74, 539–563.

Angrist, Joshua D., and William N. Evans (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 88, 450–477.

——— (1999): "Schooling and Labor Market Consequences of the 1970 State Abortion Reforms," in *Research in Labor Economics*, ed. by Solomon W. Polachek, vol. 18, pp. 75–113. Elsevier Science, Amsterdam.

Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67, 499–527.

Angrist, Joshua D., and Jinyong Hahn (2004): "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *Review of Economics and Statistics*, 86, 58–72.

Angrist, Joshua D., Guido Imbens, and Donald B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–472.

Angrist, Joshua D., and Guido W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 430–442.

Angrist, Joshua D., and Alan B. Krueger (1991): "Does Compulsory Schooling Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, 106, 976–1014.

——— (1992): "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, 418, 328–36.

——— (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business and Economic Statistics*, 13, 225–35.

——— (1999): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, ed. by Orley C. Ashenfelter, and David Card, vol. 3. North Holland, Amsterdam.

——— (2001): "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, 15, 69–85.

Angrist, Joshua D., and Guido Kuersteiner (2004): "Semiparametric Causality Tests Using the Policy Propensity Score," National Bureau of Economic Research, Working Paper No. 10975.

Angrist, Joshua D., and Kevin Lang (2004): "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program," *The American Economic Review*, 94, 1613–1634.

Angrist, Joshua D., and Victor Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533–575.

——— (2007): "The Effects of High Stakes High School Achievement Awards: Evidence from a Group-Randomized Trial," unpublished paper, Department of Economics, Massachusetts Institute of Technology.

Angrist, Joshua D., Victor Lavy, and Analia Schlosser (2006): "Multiple Experiments for the Causal Link Between the Quantity and Quality of Children," MIT Department of Economics Working Paper No. 06-26.

Arellano, Manuel, and Stephen Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *The Review of Economic Studies*, 58, 277–297.

Ashenfelter, Orley A. (1978): "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47–57.

——— (1991): "How Convincing is the Evidence Linking Education and Income?," Princeton University, Industrial Relations Section, Working Paper No. 292.

Ashenfelter, Orley A., and David Card (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *The Review of Economics and Statistics*, 67, 648–660.

Ashenfelter, Orley A., and Alan B. Krueger (1994): "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review*, 84, 1157–1173.

Ashenfelter, Orley A., and Cecilia Rouse (1998): "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins," *Quarterly Journal of Economics*, 113, 253– 284.

Athey, Susan, and Guido Imbens (2006): "Identification and Inference in Nonlinear Difference-in-Difference Models," *Econometrica*, 74, 431–497.

Atkinson, Anthony B. (1970): "On the Measurement of Inequality," *Journal of Economic Theory*, 2, 244–263.

Autor, David (2003): "Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing," *Journal of Labor Economics*, 21, 1–42.

Autor, David, Lawrence F. Katz, and Melissa S. Kearney (2005): "Rising Wage Inequality: The Role of Composition and Prices," National Bureau of Economic Research, Working Paper No. 11628.

Barnett, Steven W. (1992): "Benefits of Compensatory Preschool Education," *Journal of Human Resources*, 27, 279–312.

Barnow, Burt S., Glen G. Cain, and Arthur Goldberger (1981): "Selection on Observables," *Evaluation Studies Review Annual*, 5, 43–59.

Bekker, Paul A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62, 657–681.

Bell, Robert M., and Daniel F. McCaffrey (2002): "Bias Reduction in Standard Errors for Linear Regression with Multistage Samples," *Survey Methodology*, 28, 169–181.

Bennedsen, Morten, Kasper M. Nielsen, Francisco Pérez-González, and Daniel Wolfenzon (2007): "Inside the Family Firm: The Role of Families in Succession Decisions and Performance," *The Quarterly Journal of Economics*, 122, 647–692.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004): "How Much Should We Trust Differences-in-Differences Estimates?," *Quarterly Journal of Economics*, 119, 249–275.

Bertrand, Marianne, and Sendhil Mullainathan (2004): "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *The American Economic Review*, 94, 991–1013.

Besley, Timothy, and Robin Burgess (2004): "Can Labour Market Regulation Hinder Economic Performance? Evidence from India," *Quarterly Journal of Economics*, 113, 91–134.

Bjorklund, Anders, and Markus Jantti (1997): "Intergenerational Income Mobility in Sweden Compared to the United States," *The American Economic Review*, 87, 1009–1018.

Black, Dan A., Jeffrey A. Smith, Mark C. Berger, and Brett J. Noel (2003): "Is the Threat of Reemployment Services More Effective than the Services Themselves? Evidence from Random Assignment in the UI System," *The American Economic Review*, 93, 1313–1327.

Bloom, Howard S. (1984): "Estimating the Effect of Job-Training Programs, Using Longitudinal Data: Ashenfelter's Findings Reconsidered," *The Journal of Human Resources*, 19, 544–556.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos (1997): "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study," *The Journal of Human Resources*, 32, 549–576.

Blundell, Richard, and Stephen Bond (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87, 115–143.

Borjas, George (1992): "Ethnic Capital and Intergenerational Mobility," *Quarterly Journal of Economics*, 107, 123–150.

——— (2005): *Labor Economics*, 3rd edn. McGraw-Hill/Irwin, New York.

Bound, John, David Jaeger, and Regina Baker (1995): "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Variables is Weak," *Journal of American Statistical Association*, 90, 443–450.

Bound, John, and Gary Solon (1999): "Double Trouble: On the Value of Twins-based Estimation of the Returns of Schooling," *Economics of Education Review*, 18, 169–182.

Bronars, Stephen G., and Jeff Grogger (1994): "The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment," *American Economic Review*, 84, 1141–1156.

Buchinsky, Moshe (1994): "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression," *Econometrica*, 62, 405–458.

Buse, A. (1992): "The Bias of Instrumental Variable Estimators," *Econometrica*, 60, 173–180.

Cameron, Colin, Jonah Gelbach, and Douglas L. Miller (2008): "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, forthcoming, unpublished paper, Department of Economics, The University of California at Davis.

Campbell, Donald Thomas (1969): "Reforms as Experiments," *American Psychologist*, 24, 409–429.

Campbell, Donald Thomas, and Julian C. Stanley (1963): *Experimental and Quasi-experimental Designs for Research*. Rand McNally, Chicago.

Card, David (1992): "Using Regional Variation to Measure the Effect of the Federal Minimum Wage," *Industrial and Labor Relations Review*, 46, 22–37.

——— (1995): "Earnings, Schooling and Ability Revisited," in *Research in Labor Economics*, ed. by Solomon W. Polachek, vol. 14, pp. 23–48. JAI Press, Greenwich, Connecticut.

——— (1999): "The Causal Effect of Education on Earnings," in *Handbook of Labor Economics*, ed. by Orley C. Ashenfelter, and David Card, vol. 3. North Holland, Amsterdam.

Card, David, and Alan Krueger (1994): "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772–784.

——— (2000): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply," *American Economic Review*, 90, 1397–420.

Card, David, and David S. Lee (2008): "Regression Discontinuity Inference with Specification Error," *Journal of Econometrics*, 142, 655–674.

Card, David, and Thomas Lemieux (1996): "Wage Dispersion, Returns to Skill, and Black-White Differentials," *Journal of Econometrics*, 74, 316–361.

Card, David E., and Daniel Sullivan (1988): "Measuring the Effect of Subsidized Training on Movements In and Out of Employment," *Econometrica*, 56, 497–530.

Cardell, Nicholas Scott, and Mark Myron Hopkins (1977): "Education, Income, and Ability: A Comment," *Journal of Political Economy*, 85, 211–215.

Chamberlain, Gary (1977): "Education, Income, and Ability Revisited," *Journal of Econometrics*, 5, 241–57.

——— (1978): "Omitted Variables Bias in Panel Data: Estimating the Returns to Schooling," *Annales De L'INSEE*, 30-31, 49–82.

——— (1984): "Panel Data," in *Handbook of Econometrics*, ed. by Zvi Griliches, and Michael D. Intriligator, vol. 2, pp. 1247–1318. North Holland, Amsterdam.

——— (1994): "Quantile Regression, Censoring and the Structure of Wages," in *Proceedings of the Sixth World Cogress of the Econometrics Society, Barcelona, Spain*, ed. by Christopher A. Sims, and Jean-Jacques Laffont, pp. 179–209. Cambridge University Press, New York.

Chamberlain, Gary, and Edward E. Leamer (1976): "Matrix Weighted Averages and Posterior Bounds," *Journal of the Royal Statistical Society, Series B*, 38, 73–84.

Chernozhukov, Victor, and Christian Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

——— (2007): "A Simple Approach to Heteroskedasticity and Autocorrelation Robust Inference with Weak Instruments," unpublished paper, Department of Economics, Massachusetts Institute of Technology.

Chesher, Andrew, and Gerald Austin (1991): "The Finite-Sample Distributions of Heteroskedasticity Robust Wald Statistics," *Journal of Econometrics*, 47, 153–173.

Chesher, Andrew, and Ian Jewitt (1987): "The Bias of the Heteroskedasticity Consistent Covariance Estimator," *Econometrica*, 55, 1217–1222.

Cochran, William G. (1965): "The Planning of Observational Studies of Human Populations," *Journal of the Royal Statistical Society, Series A*, 128, 234–65.

Cook, Thomas D. (2008): "Waiting for Life to Arrive: A History of the Regression-Discontinuity Design in Psychology, Statistics, and Economics," *Journal of Econometrics*, 142, 636–654, forthcoming.

Cook, Thomas D., and Vivian C. Wong (2008): "Empirical Tests of the Validity of the Regression-Discontinuity Design," *Annales d'Economie et de Statistique*, forthcoming.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik (2006): "Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand," National Bureau of Economic Research, Technical Working Paper No. 330.

Cruz, Luiz M., and Marcelo J. Moreira (2005): "On the Validity of Econometric Techniques with Weak Instruments: Inference on Returns to Education Using Compulsory School Attendance Laws," *Journal of Human Resources*, 40, 393–410.

Currie, Janet, and Aaron Yelowitz (2000): "Are Public Housing Projects Good for Kids?," *Journal of Public Economics*, 75, 99–124.

Davidon, Russell, and James G. MacKinnon (1993): *Estimation and Inference in Econometrics*. Oxford University Press, New York.

Dearden, Lorraine, Sue Middleton, Sue Maguire, Karl Ashworth, Kate Legge, Tracey Allen, Kim Perrin, Erich Battistin, Carl Emmerson, Emla Fitzsimons, and Costas Meghir (2004): "The Evaluation of Education Maintenance Allowance Pilots: Three Years' Evidence. A Quantitative Evaluation," *Department for Education and Skills, Research Report No. 499*.

Deaton, Angus (1997): *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Johns Hopkins University Press for the World Bank, Baltimore, MD.

Dee, Thomas S., and William N. Evans (2003): "Teen Drinking and Educational Attainment: Evidence from Two-Sample Instrumental Variables Estimates," *Journal of Labor Economics*, 21, 178–209.

DeGroot, Morris H., and Mark J. Schervish (2001): *Probability and Statistics*, 3rd edn. Addison-Wesley, Boston.

Dehejia, Rajeev H. (2005): "Practical Propensity Score Matching: A Reply to Smith and Todd," *Journal of Econometrics*, 125, 355–364.

Dehejia, Rajeev H., and Sadek Wahba (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–62.

Donald, Stephen G., and Kevin Lang (2007): "Inference with Difference-in-Differences and Other Panel Data," *Review of Economics and Statistics*, 89, 221–233.

Duan, Naihua, Willard D. Manning, Jr., Carl N. Morris, and Joseph P. Newhouse (1983): "A Comparison of Alternative Models for the Models for the Demand for Medical Care," *Journal of Business & Economic Statistics*, 1, 115–126.

——— (1984): "Choosing Between the Sample-Selection Model and the Multi-Part Model," *Journal of Business & Economic Statistics*, 2, 283–289.

Durbin, James (1954): "Errors in Variables," *Review of the International Statistical Institute*, 22, 23–32.

Eicker, Friedhelm (1967): "Limit Theorems for Regressions with Unequal and Dependent Errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 59–82. University of California Press, Berkeley.

Elder, Todd E., and Darren H. Lubotsky (2008): "Kindergarten Entrance Age and Children's Achievement: Impacts of State Policies, Family Background, and Peers," *Journal of Human Resources*, forthcoming, forthcoming.

Finn, Jeremy D., and Charles M. Achilles (1990): "Answers and Questions About Class Size: A Statewide Experiment," *American Educational Research Journal*, 28, 557–77.

Firpo, Sergio (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75, 259–276.

Flores-Lagunes, Alfonso (2007): "Finite Sample Evidence of IV Estimators under Weak Instruments," *Journal of Applied Econometrics*, 22, 677–694.

Freedman, David (2005): "Linear Statistical Models for Causation: A Critical Review," in *The Wiley Encyclopedia of Statistics in Behavioral Science*, ed. by B. Everitt, and D. Howell. John Wiley, Chichester, UK.

Freeman, Richard (1984): "Longitudinal Analyses of the Effect of Trade Unions," *Journal of Labor Economics*, 3, 1–26.

Frisch, Ragnar, and Frederick V. Waugh (1933): "Partial Time Regression as Compared with Individual Trends," *Econometrica*, 1, 387–401.

Frolich, Markus, and Blaise Melly (2007): "Unconditional Quantile Treatment Effects Under Endogeneity," Centre for Microdata Methods and Practice, Working Paper No. CWP32/07.

Fryer, Roland G., and Steven D. Levitt (2004): "The Causes and Consequences of Distinctively Black Names," *The Quarterly Journal of Economics*, 119, 767–805.

Goldberger, Arthur S. (1972): "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations," University of Wisconsin, Department of Economics, Working Paper.

——— (1991): *A Course in Econometrics*. Harvard University Press, Cambridge, MA.

Gosling, Amanda, Stephen Machin, and Costas Meghir (2000): "The Changing Distribution of Male Wages in the U.K.," *Review of Economic Studies*, 67, 635–66.

Granger, Clive W. J. (1969): "Investigating Causal Relation by Econometric and Cross-Sectional Method," *Econometrica*, 37, 424–438.

Griliches, Zvi (1977): "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45, 1–22.

Griliches, Zvi, and Jerry A. Hausman (1986): "Errors in variables in panel data," *Journal of Econometrics*, 31, 93–118.

Griliches, Zvi, and William M. Mason (1972): "Education, Income, and Ability," *Journal of Political Economy*, 80, S74–S103.

Grumbach, Kevin, Dennis Keane, and Andrew Bindman (1993): "Primary Care and Public Emergency Department Overcrowding," *American Journal of Public Health*, 83, 372–378.

Guryan, Jonathan (2004): "Desegregation and Black Dropout Rates," *American Economic Review*, 94, 919–943.

Haavelmo, Trygve (1944): "The Probability Approach in Econometrics," *Econometrica*, 12, S1–S115.

Hahn, Jinyong (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–31.

Hahn, Jinyong, Petra Todd, and Wilbur van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209.

Hansen, Christian B. (2007a): "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large," *Journal of Econometrics*, 141, 597–620.

——— (2007b): "Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects," *Journal of Econometrics*, 140, 670–694.

Hansen, Lars Peter (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hausman, Jerry (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

——— (1983): "Specification and Estimation of Simultaneous Equation Models," in *Handbook of Econometrics*, ed. by Zvi Griliches, and Michael Intriligator, vol. 1, pp. 391–448. North Holland, Amsterdam.

——— (2001): "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *Journal of Econometric Perspectives*, 15, 57–67.

Hausman, Jerry, Whitney Newey, and Tiemen Wouterson (2006): "IV Estimation with Heteroskedasticity and Many Instruments," unpublished paper, Department of Economics, Massachusetts Institute of Technology.

Hay, Joel W., and Randall J. Olsen (1984): "Let Them Eat Cake: A Note on Comparing Alternative Models of the Demand for Medical Care," *Journal of Business & Economic Statistics*, 2, 279–282.

Heckman, James J. (1978): "Dummy Endogenous Variables in a Simultaneous Equations System," *Econometrica*, 46, 695–712.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd (1998): "Matching as as Econometric Evaluation Estimator," *Review of Economic Studies*, 62, 261–94.

Heckman, James J., Jeffrey Smith, and Nancy Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64, 487–535.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–89.

Hirano, Keisuke, Guido W. Imbens, Donald B. Rubin, and Xiao-Hua Zhou (2000): "Assessing the Effect of an Influenza Vaccine in an Encouragement Design," *Biostatistics*, 1, 69–88.

Hoaglin, David C., and Roy E. Welsch (1978): "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17–22.

Holland, Paul W. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–70.

Holtz-Eakin, Douglas, Whitney Newey, and Harvey S. Rosen (1988): "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56, 1371–1395.

Horowitz, Joel L. (1997): "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in *Advances in Economics and Econometrics: Theory and Applications*, ed. by David M. Kreps, and Kenneth F. Wallis, vol. 3, pp. 188–222. Cambridge University Press, Cambridge.

——— (2001): "The Bootstrap," in *Handbook of Econometrics*, ed. by James J. Heckman, and Edward E. Leamer, vol. 5, pp. 3159–3228. Elsevier Science, Amsterdam.

Horvitz, Daniel G., and Donovan J. Thompson (1952): "A Generalization of Sampling Without Replacement From a Finite Population," *Journal of the American Statistical Association*, 47, 663–85.

Hoxby, Caroline (2000): "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *The Quarterly Journal of Economics*, 115, 1239–1285.

Hsia, Judith, Robert D. Langer, JoAnn E. Manson, Lewis Kuller, Karen C. Johnson, Susan L. Hendrix, Mary Pettinger, Susan R. Heckbert, Nancy Greep, Sybil Crawford, Charles B. Eaton, John B. Kostis, Pat Caralis, Ross Prentice, and for the Women's Health Initiative Investigators (2006): "Conjugated Equine Estrogens and Coronary Heart Disease: The Women's Health Initiative," *Archives of Internal Medicine*, 166, 357–365.

Imbens, Guido (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706–10.

——— (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

Imbens, Guido, and Joshua Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–476.

Imbens, Guido, and Thomas Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615–635.

Inoue, Atsushi, and Gary Solon (2005): "Two-Sample Instrumental Variables Estimators," National Bureau of Economic Research, Technical Working Paper No. 311.

Jappelli, Tullio, Jorn-Steffen Pischke, and Nicholas S. Souleles (1998): "Testing for Liquidity Constraints in Euler Equations with Complementary Data Sources," *The Review of Economics and Statistics*, 80, 251–262.

Johnson, Norman L., and Samuel Kotz (1970): *Distributions in Statistics: Continuous Distributions*, vol. 2. John Wiley, New York.

Kauermann, Goran, and Raymond J. Carroll (2001): "A Note on the Efficiency of Sandwich Covariance Estimation," *Journal of the American Statistical Association*, 96, 1387–1396.

Kelejian, Harry H. (1971): "Two Stage Least Squares and Econometric Systems Linear in Parameters but Non-linear in the Endogenous Variables," *Journal of the American Statistical Association*, 69, 373–4.

Kennan, John (1995): "The Elusive Effects of Minimum Wages," *Journal of Economic Literature*, 33, 1950–1965.

Kézdi, Gábor (2004): "Robust Standard Error Estimation in Fixed-Effects Panel Models," *Hungarian Statistical Review, Special English Volume*, 9, 95–116.

Kish, Leslie (1965): "Sampling Organizations and Groups of Unequal Sizes," *American Sociological Review*, 30, 564–572.

Kloek, Teun (1981): "OLS Estimation in a Model Where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated," *Econometrica*, 49, 205–207.

Knight, Keith (2000): *Mathematical Statistics*. Chapman & Hall/CRC, Boca Raton, FL.

Koenker, Roger (2005): *Quantile Regression*. Cambridge University Press.

Koenker, Roger, and Gilbert Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, Roger, and Stephen Portnoy (1996): "Quantile Regression," University of Illinois at Urbana-Champaign, College of Commerce and Business Administration, Office of Research, Working Paper No. 97-0100.

Krueger, Alan B. (1999): "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114, 497–532.

Kugler, Adriana, Juan F. Jimeno, and Virginia Hernanz (2005): "Employment Consequences of Restrictive Permanent Contracts: Evidence from Spanish Labor Market Reforms," FEDEA Working Paper 2003-14.

LaLonde, Robert (1995): "The Promise of Public Sector-Sponsored Training Programs," *Journal of Economic Perspectives*, 93, 149–68.

LaLonde, Robert J. (1986): "Evaluating the Econometric Evaluations of Training Programs Using Experimental Data," *American Economic Review*, 76, 602–20.

Lee, David S. (2008): "Randomized experiments from non-random selection in U.S. House elections," *Journal of Econometrics*, 142, 675–697.

Lemieux, Thomas (2008): "The Changing Nature of Wage Inequality," *Journal of Population Economics*, 21, 21–48.

Liang, Kung-Yee, and Scott L. Zeger (1986): "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Machado, Jose, and Jose Mata (2005): "Counterfactual Decompositions of Changes in Wage Distributions Using Quantile Regression," *Journal of Applied Econometrics*, 20, 445–65.

MacKinnon, James G., and Halbert White (1985): "Some Heteroskedasticity Consistent Covariance Matrix Estimators With Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305–325.

Maddala, Gangadharrao Soundalyarao (1983): "Methods of Estimation for Models of Markets with Bounded Price Variation," *International Economic Review*, 24, 361–378.

Mammen, Enno (1993): "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *Annals of Statistics*, 21, 255–285.

Manning, Willard G., Joseph P. Newhouse, Naihua Duan, Emmett B. Keeler, Arleen Leibowitz, and Susan M. Marquis (1987): "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *American Economic Review*, 77, 251–77.

Manski, Charles F. (1991): "Regression," *Journal of Economic Literature*, 29, 34–50.

Mariano, Roberto S. (2001): "Simultaneous Equation Model Estimators: Statistical Properties," in *A Companion to Theoretical Econometrics*, ed. by B. Baltagi. Blackwell, Oxford.

McClellan, Mark B., Barbara J. McNeil, and Joseph P. Newhouse (1994): "Does More Intensive Treatment of Acute Myocardial Infarction Reduce Mortality? Analysis Using Instrumental Variables," *Journal of the American Medical Association*, 272, 859–66.

McCrary, Justin (2008): "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142, 698–714.

McDonald, John F., and Robert A. Moffitt (1980): "The Uses of Tobit Analysis," *The Review of Economics and Statistics*, 62, 318–321.

Melly, Blaise (2005): "Decomposition of Differences in Distribution Using Quantile Regression," *Labour Economics*, 12, 577–590.

Meltzer, Allan H., and Scott F. Richard (1983): "Tests of a Rational Theory of the Size of Government," *Public Choice*, 41, 403–418.

Messer, Karen, and Halbert White (1984): "A Note on Computing the Heteroskedasticity Consistent Covariance Matrix Using Instrumental Variables Techniques," *Oxford Bulletin of Economics and Statistics*, 46, 181–184.

Meyer, Bruce, Kip Viscusi, and David Durbin (1995): "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review*, 85, 322–340.

Meyer, Bruce D., and Dan T. Rosenbaum (2001): "Welfare, the Earned Income Tax Credit, and the Labor Supply of Single Mothers," *The Quarterly Journal of Economics*, 116, 1063–1114.

Milgram, Stanley (1963): "Behavioral Study of Obedience," *Journal of Abnormal and Social Psychology*, 67, 371–78.

Moffitt, Robert (1992): "Incentive Effects of the U.S. Welfare System: A Review," *Journal of Economic Literature*, 30, 1–61.

Morgan, Mary S. (1990): *The History of Econometric Ideas*. Cambridge University Press, Cambridge, U.K.

Moulton, Brent (1986): "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, 32, 385–397.

Nelson, Charles R., and Richard Startz (1990a): "The Distribution of the Instrumental Variables Estimator and Its t-ratio when the Instrument is a Poor One," *Journal of Business*, 63, 125–140.

——— (1990b): "Some Further Results on the Exact Small-Sample Properties of the Instrumental Variable Estimator," *Econometrica*, 58, 967–976.

Neumark, David, and William Wascher (1992): "Employment Effects of the Minimum and Subminimum Wages: Panel Data on State Minimum Wage Laws," *Industrial and Labor Relations Review*, 46, 55–81.

Newey, Whitney K. (1985): "Generalized Method of Moments Specification Testing," *Journal of Econometrics*, 29, 299–256.

——— (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

Newey, Whitney K., and Kenneth D. West (1987): "Hypothesis Testing with Efficient Method of Moments Estimation," *International Economic Review*, 28, 777–787.

Nickell, Stephen (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49, 1417–1426.

Obenauer, Marie, and Berta von der Nienburg (1915): "Effect of Minimum Wage Determinations in Oregon," Bulletin of the US Bureau of Labor Statistics, No. 176, US GPO, Washington, D.C.

Oreopoulos, Philip (2006): "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *American Economic Review*, 96, 152–175.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Fred Doolittle, and Winston Lin (1996): *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study*. Urban Institute Press, Washington, D.C.

Pfefferman, Daniel (1993): "The Role of Sampling Weights When Modeling Survey Data," *International Statistical Review*, 61, 317–37.

Pischke, Jorn-Steffen (2007): "The Impact of Length of the School Year on Student Performance and Earnings: Evidence from the German Short School Years," *Economic Journal*, 117, 1216–1242.

Porter, Jack (2003): "Estimation in the Regression Discontinuity Model," unpublished paper, Department of Economics, Harvard University.

Poterba, James, Steven Venti, and David Wise (1995): "Do 401K Contributions Crowd Out Other Personal Savings," *Journal of Public Economics*, 58, 1–32.

Powell, James L. (1986): "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155.

——— (1989): "Semiparametric Estimation of Censored Selection Models," unpublished paper, Department of Economics, University of Wisconsin-Madison.

Prais, Sig J., and John Aitchison (1954): "The Grouping of Observations in Regression Analysis," *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 22, 1–22.

Reiersol, Olav (1941): "Confluence Analysis by Means of Lag Moments and Other Methods of Confluence Analysis," *Econometrica*, 9, 1–24.

Robins, James M., Steven D. Mark, and Whitney K. Newey (1992): "Estimating Exposure Effects by Modeling the Expectation of Exposure Conditional on Confounders," *Biometrics*, 48, 479–95.

Rosenbaum, Paul R. (1984): "The Consequences of Adjustment for a Concomitant Variables that has Been Affected by the Treatment," *Journal of the Royal Statistical Society, Series A*, 147, 656–66.

——— (1995): *Observational Studies*. Springer-Verlag, New York.

Rosenbaum, Paul R., and Donald B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

——— (1985): "The Bias Due to Incomplete Matching," *Biometrics*, 41, 106–116.

Rosenzweig, Mark R., and Kenneth I. Wolpin (1980): "Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment," *Econometrica*, 48, 227–240.

Rubin, Donald B. (1973): "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159–83.

——— (1974): "Estimating the Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1977): "Assignment to a Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26.

——— (1991): "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism," *Biometrics*, 47, 1213–34.

Ruud, Paul A. (1986): "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution," *Journal of Econometrics*, 32, 157–87.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell (2002): *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton-Mifflin Company, Boston.

Sherman, Lawrence W., and Richard A. Berk (1984): "The Specific Deterrent Effects of Arrest for Domestic Assault," *American Sociological Review*, 49, 261–272.

Shore-Sheppard, Lara (1996): "The Precision of Instrumental Variables Estimates with Grouped Data," Princeton University, Industrial Relations Section, Working Paper No. 374.

Smith, Jeffrey A., and Petra E. Todd (2001): "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," *American Economic Review*, 91, 112–118.

——— (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, 125, 305–53.

Stigler, Stephen M. (1986): *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press, Cambridge, MA.

Stock, James H., and Francesco Trebbi (2003): "Who Invented Instrumental Variables Regression?," *The Journal of Economic Perspectives*, 17, 177–94.

Stock, James H., Jonathan H. Wright, and Motohiro Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20, 518–529.

Taubman, Paul (1976): "The Determinants of Earnings: Genetics, Family and Other Environments: A Study of White Male Twins," *American Economic Review*, 66, 858–870.

Thistlewaite, Donald L., and Donald T. Campbell (1960): "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, 51, 309–317.

Trochim, William (1984): *Research Designs for Program Evaluation: The Regression Discontinuity Design*. Sage Publications, Beverly Hills, CA.

van der Klaauw, Wilbert (2002): "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach," *International Economic Review*, 43.

Wald, Abraham (1940): "The Fitting of Straight Lines if Both Variables are Subject to Error," *Annals of Mathematical Statistics*, 11, 284–200.

——— (1943): "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," *Transactions of the American Mathematical Society*, 54, 426–482.

White, Halbert (1980a): "A Heteroskedasticity-Consistent Covariance Matric Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–38.

——— (1980b): "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21, 149–170.

——— (1982): "Instrumental Variables Regression with Independent Observations," *Econometrica*, 50, 483–499.

Wooldridge, Jeffrey (2003): "Cluster-sample Methods in Applied Econometrics," *American Economic Review*, 93, 133.

——— (2005): "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *The Review of Economics and Statistics*, 87, 385–390.

——— (2006): *Introductory Econometrics: A Modern Approach*. Thomson/South-Western, Mason, OH.

Wright, Phillip G. (1928): *The Tariff on Animal and Vegetable Oils*. Macmillan, New York.

Yang, Song, Li Hsu, and Lueping Zhao (2005): "Combining Asymptotically Normal Tests: Case Studies in Comparison of Two Groups," *Journal of Statistical Planning and Inference*, 133, 139–158.

Yelowitz, Aaron (1995): "The Medicaid Notch, Labor Supply and Welfare Participation: Evidence from Eligibility Expansions," *The Quarterly Journal of Economics*, 110, 909–939.

Yule, George Udny (1895): "On the Correlation of Pauperism with Proportion of Out-Relief," *The Economic Journal*, 5, 603–611.

——— (1897): "On the Theory of Correlation," *Journal of the Royal Statistical Society*, 60, 812–854.

——— (1899): "An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades (Part I.)," *Journal of the Royal Statistical Society*, 62, 249–295.